Evolution du régulateur floral LEAFY dans la lignée verte

Marie Monniaux

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Biologie végétale**

Arrêté ministériel : 7 août 2006

Présentée par

Marie MONNIAUX

Thèse dirigée par **François PARCY**

préparée au sein du **Laboratoire de Physiologie Cellulaire et Végétale**

dans **l'École Doctorale Chimie et Sciences du Vivant**

Evolution du régulateur floral LEAFY dans la lignée verte

Thèse soutenue publiquement le **11 Décembre 2012**,
devant le jury composé de :

Pr. Christian FANKHAUSER
Rapporteur
Dr. Françoise MONEGER
Rapporteur
Pr. Christophe ROBAGLIA
Examinateur
Dr. Marc BILLAUD
Examinateur
Dr. François PARCY
Directeur de thèse

Remerciements

Je remercie tout d'abord Norbert Rolland, Laurent Blanchoin et Marylin Vantard pour m'avoir accueillie au sein du laboratoire. Je remercie également les membres de mon jury pour avoir accepté d'évaluer ce travail de thèse.

Le plus grand merci revient bien sûr à François. Malgré tes nombreuses occupations, tu as toujours été très disponible et tu as toujours pris le temps de discuter avec moi de mes derniers résultats, ou de choses et d'autres. Et même si tu es loin de Grenoble pour un an, je sais bien que tu seras toujours très attentif à ce qu'il se passe au sein de l'équipe. Merci pour ce formidable enthousiasme scientifique, pour la rigueur que tu as pu me transmettre (j'en avais bien besoin !), j'espère pouvoir conserver toutes ces choses précieuses pour plus tard. J'espère trouver une équipe qui tourne aussi bien par la suite, mais je sais que sans le moteur "François", ça ne sera peut-être pas aussi bien ! Et merci bien sûr pour toutes tes histoires de courses, de montagne, de famille et de science que tu partages avec nous, je crois que tes histoires à la cantine, c'est ce qui me manquera le plus cette année !

Un merci particulier à Edwige, avec qui j'ai beaucoup travaillé sur les SELEX et autres, et qui a un peu représenté mon modèle de thésarde pendant ma première année ! Tu m'as beaucoup impressionnée par ta passion pour LFY et les fleurs, et ton implication dans ton travail. C'est toujours une joie de te revoir au labo, et j'espère que nos chemins se recroiseront de temps en temps !

Plein de mercis au lab manager sans qui rien de tout ça n'aurait été possible : Manou ! C'est toi qui m'as montré mon premier gel retard, quand on sait combien j'en ai fait après... Heureusement que tu es là, sinon l'équipe (voire même le labo) s'effondrerait ! Merci aussi pour les randos, pour l'Ekiden (oui oui même ça !) et pour tous les « gentils » surnoms que tu m'as donnés :)

Un grand merci à Camille, pour ton calme et ta bonne humeur, c'est un vrai plaisir de travailler avec toi. J'essaye de prendre exemple sur ta patience et ta rigueur quand tu manipes, j'ai encore des progrès à faire si je veux arriver à ton niveau ! Merci aussi pour toute la musique qui a rythmé nos manips. Et surtout n'oublies pas, on peut purifier mille protéines mille fois, mais on ne peut pas purifier une protéine mille fois... (?!?)

Merci à Hichaminator le bibliovore, un grand passionné de LFY ! J'ai l'impression que tu es dans le labo depuis 10 ans tellement tu connais les publis sur le bout des doigts, la relève est assurée. Merci aussi pour ton aide pour les plantes, et pour notre initiation au geocaching et à Just Dance !

Merci à mon nouveau co-bureau Renaud, avec qui c'est toujours si drôle de travailler. C'est super d'avoir toujours de bons conseils de biochimie à portée de main, et quelqu'un toujours partant pour une pause café aussi ! Tu as toujours la petite phrase pour faire rire :) Merci aussi pour tes corrections de ce manuscrit. Une petite pensée pour Michel également, notre presque troisième co-bureau ;)

Pour les pauses café, je compte toujours aussi sur Gaby ! Merci pour les discussions de science, d'enseignement, et de tout et de rien, c'est toujours un plaisir. Merci à la "grande Marie" qu'on espère revoir très bientôt, même si on comprend bien que le petit Noa passe avant tout! Merci aussi à Gilles, notre maître yogi à tous.

Fanny, merci pour les grands éclats de rire qui traversent le labo ! La ULT-team est entre de bonnes mains. Un très grand merci à Cristel pour ta gentillesse, ton aide pour les in situs et les enseignements. C'est chouette de te voir si motivée par la science ! Merci à Robert également, pour les discussions sur tout et rien, le Python, la course ou la Sibérie ! Thanks to our new recruit Julia, it has been very nice to chat with you while I was writing, and I am happy that we still have another year to share in the team !

Merci aussi aux anciens de l'équipe: Sandrine, qu'on aurait bien aimé avoir à nouveau dans l'équipe ! J'espère que tu trouveras très vite une nouvelle chaussure à ton pied. Une pensée pour Eugenio, pour tes super compétences en programmation, et aussi pour tes très nombreux gâteaux ! Thanks to Elena, it has been a real pleasure to work with you for a few months and I hope you will find the mystery of STOP1...

Merci à nos collaborateurs, Fabien et Florence, qui m'ont patiemment appris à faire pousser des p'tites mousses, et toute l'équipe de Mohammed pour nos histoires de roses. Merci également à Téva et Dominique pour m'avoir donné de très bons conseils au cours de ma thèse.

Un énorme merci à tous les jeunes et moins jeunes du labo, les anciens et les nouveaux: Matthieu, Florie, Thomas, Morgane, Cécile, Cristian, Elisa, Martino, Jim, Didier, Timothée, Laurence, Fabien, Guillaume, Cécile, et tous les membres de PCV qui font de ce labo un endroit où on est simplement contents de venir tous les matins. Merci à Lucas et Daniel pour une gestion impeccable du stock des enzymes :) Merci à Sassia pour tes milieux et pour ta gentillesse. Un très grand merci à Sophie et Sylvianne, pour leur bonne humeur et leur super-efficacité, même quand je n'ai pas rempli d'ordre de mission avant de partir !

Merci aux [BioGeeks] pour les loooongs échanges de mails qui montrent que finalement, médecins, thésards ou profs, on est tous dans la même galère ! Merci à Charlotte et à Damien pour les week-end sympas entre « lyonnais », et merci à Elise, Angélique et Véro pour les (trop rares) après-midis thé-biscuits ! Un merci spécial à Edouard, que j'aimerais bien pouvoir voir plus souvent.

Merci à papa et maman qui m'ont toujours encouragée dans cette voie. Merci à mamie aussi, et une grosse pensée pour papy qui aurait été content de voir sa petite-fille soutenir sa thèse, j'en suis sûre. Merci à Elise, Mohammed, Sylvain et Caroline (et même au petit Gabichon !) sur qui je pourrai toujours compter.
Pour Ronan, un simple merci ne suffira jamais pour tout ce que tu m'as apporté.

SOMMAIRE

ABREVIATIONS 5

INTRODUCTION 6

I) Les régulations transcriptionnelles : prédiction et évolution 7

1) Qu'est-ce qu'un réseau de régulation ? 7

2) Evolution d'une régulation : éléments cis ou trans ? 8

3) Prédire une régulation transcriptionnelle et son évolution 11

II) LFY et le réseau floral chez les angiospermes 14

1) Présentation du facteur de transcription LFY chez Arabidopsis thaliana 14

2) Le réseau floral contrôlé par LFY 17

III) Evolution de LFY et de son réseau chez les plantes terrestres 19

1) LFY chez les plantes « sans fleurs » 19

2) Un rôle ancestral pour LFY ? 21

3) Evolution de LFY et de son réseau 22

Article 1 25

PRESENTATION DES OBJECTIFS 32

RESULTATS 34

CHAPITRE I : Evolution de la spécificité de liaison à l'ADN de LFY chez les plantes terrestres 34

1) Production de protéines LFY recombinantes 35

2) Expériences de SELEX 37

3) Séquençage haut-débit des échantillons de SELEX 39

4) Analyse des résultats de SELEX : spécificité de liaison à l'ADN de LFY chez les plantes terrestres 41

CHAPITRE II : Prédiction de la liaison de LFY à ses gènes cibles 45

1) Prédire la liaison de LFY à l'ADN 46

2) Prédiction de la régulation des gènes floraux par LFY chez les angiospermes et les gymnospermes 49

2) Des prédictions à l'échelle génomique ? 55

Article 2 ..... *59*
Article 3 ..... *74*
Article 4 ..... *88*

CHAPITRE III : Etude du changement de spécificité de LFY chez *Physcomitrella patens* 95
1) Optimisation de la matrice de PpLFY1 in vitro *96*
2) Prédiction des sites de liaison de PpLFY1 dans le génome de P. patens *99*
3) Recherche des gènes régulés par PpLFY1 ..... *100*
4) Evolution moléculaire du changement de spécificité LFY-PpLFY1 ..... *104*

DISCUSSION ET PERSPECTIVES 112

I) LFY et son évolution dans la lignée verte : une histoire originale et variée 112
1) Un gène unique *112*
2) Evolution cis ou trans ? *115*
II) Prédire des gènes cibles et une fonction ancestrale pour LFY ? 119
1) Prédire une liaison, prédire une régulation *119*
2) La fonction ancestrale de LFY : des prédictions aux expérimentations *123*

MATERIEL ET METHODES 126

I) Détermination de la spécificité de liaison à l'ADN de LFY : des clonages au SELEX 126
II) Prédire la liaison de LFY à l'ADN 129
III) Culture de *Physcomitrella patens*, et méthodes de biologie moléculaire associées ... 130

REFERENCES BIBLIOGRAPHIQUES 132

ABREVIATIONS

35S : promoteur constitutif fort en amont de l'ARN 35S du virus de la mosaïque du chou-fleur (CaMV)
ADN : Acide DesoxyriboNucléique
ADNc : ADN complémentaire, issu de la rétrotranscription d'un ARN messager
AG : *AGAMOUS*, gène cible de LFY participant à l'identité des organes reproducteurs de la fleur
AP1 : *APETALA1*, gène cible de LFY
ChIP-Seq et ChIP-chip : Chromatin ImmunoPrecipitation (Immunoprécipitation de la chromatine) suivie d'un séquençage massif (Seq) ou de l'hybridation des séquences sur puce (chip)
DNase (ADNase): DesoxyriboNucléase, enzyme de dégradation de l'ADN
EMSA : Electrophoretic Mobility Shift Assay, ou gel retard
GO : Gene Ontology
K_D : Constante de dissociation
kDa : kiloDalton
LFY : LEAFY
LFY-C : domaine C-terminal de LFY
LFYΔ : protéine LFY tronquée d'environ 40 acides aminés du côté N-terminal
MADS-box : domaine des gènes homéotiques de type MADS, du nom des premiers membres identifiés (MCM1, AGAMOUS, DEFICIENS, SRF)
MEME : Multiple EM for Motif Elicitation, programme informatique identifiant un motif surreprésenté dans un jeu de séquences
MF : Méristème Floral, structure indifférenciée initiée par le méristème d'inflorescence, et développant les organes floraux
MI : Méristème d'Inflorescence, structure indifférenciée au sommet de la tige des angiospermes après la transition florale, initiant des méristèmes floraux sur ses flancs
NLY : *NEEDLY*, paralogue de *LFY* chez les gymnospermes
pb : paire de bases
POcc : Predicted Occupation (Occupation prédite), reflétant le nombre de molécules d'un facteur de transcription liées à un fragment d'ADN
PpLFY1 **et** ***PpLFY2*** : *Physcomitrella patens LFY* 1 et 2
PWM : Position Weight Matrix, matrice poids/position associant un poids à chaque nucléotide du site de liaison d'un facteur de transcription
QuMFRA : Quantitative Multiple Fluorescence Relative Affinity
ROC-AUC : Receiver Operating Characteristic – Area Under the Curve, aire sous une courbe ROC, comparant le taux de vrais et de faux positifs prédits par un modèle
RT-PCR : Reverse Transcription (Rétrotranscription) suivie d'une PCR (Polymerase Chain Reaction)
SAM : Shoot Apical Meristem (Méristème Apical Caulinaire), structure indifférenciée au sommet de la tige des angiospermes, responsable de la formation des organes latéraux
SELEX : Systematic Evolution of Ligands by EXponential enrichment
SHP : *SHATTERPROOF*
VP16 : protéine du virus de l'Herpès, possédant un domaine fort d'activation de la transcription
wt : wild-type, plante sauvage

INTRODUCTION

Une des grandes questions de la biologie évolutive est de savoir comment sont apparues les morphologies actuelles des êtres vivants. Quelle est l'origine d'un œil, d'une aile ou bien d'une racine ? Comment des structures aussi complexes ont-elles pu être créées au cours de l'évolution ? Ont-elles parfois une origine commune ? L' « évo-dévo » s'est développée assez récemment en réponse à ces questions. Cette discipline recherche ainsi par quels processus développementaux sont apparues des structures morphologiques, et plus précisément quels sont les mécanismes génétiques sous-jacents qui ont permis cette apparition (Arthur, 2002; Muller, 2007).

En biologie végétale, la question de l'apparition de la fleur, décrite par Charles Darwin comme un « abominable mystère » dans une lettre au botaniste J.D. Hooker en 1879, reste encore irrésolue. Pourquoi un tel mystère ? Car la fleur présente une structure unique et originale : présence de sépales et de pétales aux fonctions protectrices et attractives pour les pollinisateurs, ovules protégés dans des carpelles, qui permettent la protection de la graine dans un fruit, regroupement des organes reproducteurs mâles et femelles sur le même axe… Ces innovations ont participé à l'immense succès évolutif des angiospermes (plantes à fleurs), qui ont recouvert la surface de la terre en seulement 140 à 180 millions d'années, se sont diversifiées très rapidement et représentent actuellement plus de 300 000 espèces (Friis et al., 2005; Soltis et al., 2008). De plus, aucun des fossiles découverts jusqu'à présent ne peuvent témoigner d'un intermédiaire convaincant entre la fleur et les structures reproductrices précédemment apparues lors de l'évolution (les cônes des gymnospermes) (Frohlich and Chase, 2007), ce qui souligne à nouveau la radiation très rapide des angiospermes.

Dans ce contexte, comprendre comment une structure telle que la fleur est apparue passe d'abord par la compréhension de son développement chez les espèces actuelles, et plus précisément du réseau génétique qui le contrôle. Ces réseaux génétiques, dont les éléments clés sont les facteurs de transcription, ont un rôle majeur dans l'évolution et l'apparition de nouvelles structures. Comprendre l'évolution des réseaux de régulation transcriptionnelle constitue donc une étape primordiale pour comprendre l'évolution des morphologies.

I) Les régulations transcriptionnelles : prédiction et évolution

1) Qu'est-ce qu'un réseau de régulation ?

Tous les processus biologiques sont contrôlés par un ensemble de gènes, qui constituent un réseau génétique. Ces réseaux sont constitués de gènes régulateurs, qui codent des facteurs de transcription, protéines capables de réguler l'expression de leurs gènes cibles. Un réseau de régulation est donc classiquement constitué de deux types d'éléments : les éléments *trans*, c'est-à-dire les facteurs de transcription ; et les éléments *cis*, qui sont les sites de liaison du facteur de transcription en amont des gènes régulés. Ainsi, un facteur de transcription peut réguler l'expression d'un nombre très variable de gènes selon le nombre d'éléments *cis* accessibles, ce qui peut varier au cours du temps et selon le tissu considéré.

Les réseaux de régulation ont une grande importance dans tous les processus biologiques majeurs. La mutation individuelle de nombreux gènes régulateurs peut conduire à des cas de létalité ou à des forts défauts développementaux chez la plante *Arabidopsis thaliana* ; c'est le cas par exemple de la mutation *mp* (gène *MONOPTEROS)*, qui engendre une plantule sans racine (Berleth and Jürgens, 1993), ou encore de la mutation *stm* (gène *SHOOT MERISTEMLESS)* sous laquelle la plantule ne peut pas développer ses organes aériens (tige, feuilles) (Barton and Poethig, 1993).

En outre, les modifications des réseaux de régulation sont un des plus importants moteurs de l'évolution. Le cas des gènes homéotiques *Hox* chez les bilatériens constitue un système très étudié depuis de nombreuses années. Les gènes *Hox* sont organisés en un ou plusieurs clusters de gènes, activés tour à tour au cours du développement de l'individu, qui codent des facteurs de transcription participant à l'identité des structures le long de l'axe antéro-postérieur du corps. Chez la souris ou le poussin, le patron d'expression spatial et temporel de chacun des gènes *Hox* est bien décrit (Wellik, 2009). On observe cependant de nombreuses variations autour de ce modèle selon les espèces étudiées. Chez le python, qui présente à la fois un très important allongement du corps et une perte de l'émergence des membres, le patron d'expression des gènes *Hoxc8* et *Hoxc6* est beaucoup plus étendu postérieurement que chez le poussin, ce qui va spécifier aux segments associés une identité thoracique (Cohn and Tickle, 1999). L'extension du patron d'expression de ces gènes entraîne la perte de l'activation d'autres gènes développementaux (*Fgf2* entre autres) à la surface du bourgeon de membre, ce qui n'autorise pas son émergence (Cohn and Tickle, 1999). La sortie du membre peut pourtant être partiellement restaurée en y greffant des billes recouvertes de

FGF2, ce qui signifie que le réseau génétique de réponse à FGF2 est toujours présent et fonctionnel chez le python, mais qu'il n'est plus activé *in vivo* (Cohn and Tickle, 1999).

Ainsi, la simple extension du domaine d'expression de gènes « maîtres » comme les gènes *Hox* peut entraîner des changements morphologiques majeurs au cours de l'évolution.

2) Evolution d'une régulation : éléments *cis* ou *trans* ?

Quels sont les processus moléculaires qui permettent l'évolution des réseaux génétiques ? Cette question fondamentale est encore sujette à discussion, puisque de nombreux cas différents ont été répertoriés. Si l'on considère une régulation transcriptionnelle simple, où un facteur de transcription, aidé d'un corégulateur, régule l'expression d'un gène en se fixant sur un élément *cis*, il existe théoriquement trois possibilités pour que cette régulation transcriptionnelle soit modifiée (**Fig. 1**) : **(1)** une mutation au niveau du facteur de transcription, qui devient incapable de reconnaître l'élément *cis* ; **(2)** une mutation au niveau de l'élément *cis*, qui n'est donc plus reconnu par le facteur de transcription ; **(3)** ou bien au niveau des corégulateurs qui ne sont plus présents ou fonctionnels pour assister le facteur de transcription dans sa fonction. La question de savoir si les mutations sont plus fréquentes en *cis* ou en *trans* est encore fortement débattue (Hoekstra and Coyne, 2007; Pennisi, 2008).

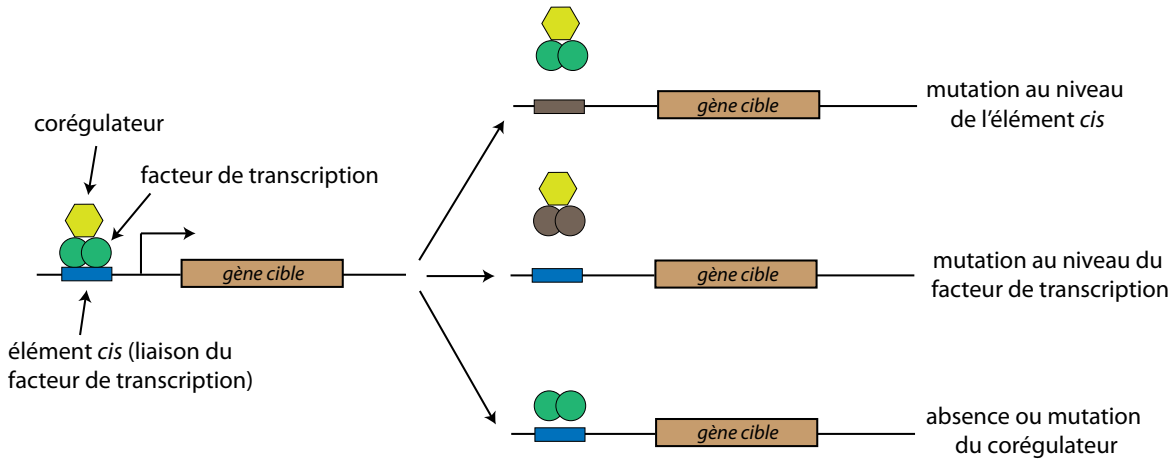


Figure 1 : Possibilités théoriques d'évolution d'une régulation transcriptionnelle. On considère une régulation impliquant un facteur de transcription, qui reconnaît un élément *cis* sur l'ADN, avec l'aide d'un corégulateur. Ces trois éléments peuvent être modifiés au cours de l'évolution, ce qui aboutit à la perte de la régulation du gène cible.

a) Evolution des éléments *trans*

Les éléments *trans* évoluent très fréquemment par le biais de duplications. En effet, des gènes ou des segments entiers du génome peuvent subir ce processus au cours de l'évolution, et les gènes régulateurs y sont particulièrement sujets (Wray et al., 2003). Ceci est

en outre très prononcé chez les plantes terrestres (Shiu et al., 2005), en partie parce qu'elles ont subi de nombreux évènements de duplication complète du génome (Adams and Wendel, 2005). Les facteurs de transcription des plantes terrestres forment donc très souvent des familles multigéniques. La famille des gènes homéotiques MADS (du nom des quatre premiers membres identifiés : *MCM1*, *AGAMOUS*, *DEFICIENS* et *SRF*), impliqués dans de nombreux processus développementaux et notamment dans la floraison, illustre bien cet exemple puisqu'on dénombre plus de 100 gènes MADS chez *Arabidopsis thaliana* (Airoldi and Davies, 2012), par opposition aux 17 représentants identifiés jusqu'à présent chez la mousse *Physcomitrella patens*, plus proche de l'ancêtre commun des plantes terrestres (Singer et al., 2007; Zobell et al., 2010).

Si la duplication est maintenue lors de l'évolution, l'un des deux gènes résultants peut évoluer de plusieurs façons. **(1)** Il peut conserver la fonction du gène ancestral, ce qui aboutit à une redondance fonctionnelle totale entre les deux gènes. Ce processus est assez courant, puisque chez la levure *Saccharomyces cerevisiae*, on dénombre 239 paires de gènes redondants (Kafri et al., 2009). **(2)** Il peut également acquérir une nouvelle fonction (néofonctionnalisation). C'est le cas des proches paralogues *AGAMOUS (AG)*, *SHATTERPROOF (SHP)* et *SEEDSTICK (STK)* (gènes MADS) chez *A. thaliana*, *AG* étant impliqué dans l'identité des organes reproducteurs (étamines et carpelles), alors que *SHP* est nécessaire à la maturation du fruit et que *STK* est impliqué dans l'abscission des graines (Pinyopich et al., 2003). **(3)** Enfin, les deux gènes (ancestral et dupliqué) peuvent se partager la fonction ancestrale (subfonctionnalisation). Chez le muflier (*Antirrhinum majus*), l'identité des organes reproducteurs est assurée par les deux gènes *FARINELLI* et *PLENA*, alors qu'elle n'est portée que par *AG* chez *A. thaliana* (Causier et al., 2005). Les trois situations ne sont pas mutuellement exclusives, puisqu'on observe souvent des cas de redondance partielle entre gènes. A nouveau, les gènes *AG*, *SHP* et *STK*, qui ont une fonction non redondante dans les processus biologiques cités précédemment, sont pourtant redondants pour définir l'identité de l'ovule (Pinyopich et al., 2003). Ainsi, le processus de duplication d'un gène régulateur apporte une énorme flexibilité au réseau de régulation, en lui permettant d'évoluer de nombreuses manières tout en gardant la fonction du gène ancestral.

L'évolution d'un facteur *trans* est-elle néanmoins possible sans duplication ? Ceci paraît difficile théoriquement, puisque modifier les propriétés (spécificité de liaison à l'ADN, activité transcriptionnelle,…) du facteur « maître » du réseau entraînerait la modification de la régulation de tous ses gènes cibles, ce qui pourrait avoir des conséquences fonctionnelles très importantes. Le cas du gène *Ubx*, qui appartient au groupe des gènes *Hox*, a néanmoins été

bien décrit chez la drosophile et le crustacé *Artemia*. Une mutation de la protéine Ubx, localisée dans son homéodomaine (domaine de liaison à l'ADN), a été identifiée chez *Artemia*, où Ubx est alors incapable de réprimer l'émergence des membres dans les segments abdominaux postérieurs, ce qui explique la différence entre les six pattes des insectes et les multiples pattes des crustacés (Ronshaugen et al., 2002). Cette mutation est structurale, concerne un facteur de transcription, et il n'y a pas de redondance connue avec un autre facteur homéotique. Il est donc possible d'avoir une mutation en *trans*, sans duplication du gène régulateur, qui n'aboutisse pas à des conséquences dramatiques pour le développement de l'individu. Néanmoins, d'après le nombre d'exemples fournis par la littérature (Hoekstra and Coyne, 2007), ceci semble assez rare.

b) Evolution des éléments *cis*

Les mutations des éléments régulateurs (éléments *cis*) sont plus souvent invoquées pour expliquer l'évolution des réseaux de régulation. En théorie, la modification d'un seul élément *cis* en amont d'un gène pourra modifier très finement et graduellement le réseau puisque seule la régulation de ce gène sera perturbée.

De nombreux cas d'évolution en *cis* sont bien documentés, comme par exemple celui de la résistance des drosophiles à l'agent insecticide DDT (dichloro-diphényl-trichloroéthane). La surexpression du gène *Cyp6g1*, codant le cytochrome P450, est nécessaire et suffisante pour qu'une drosophile acquière la résistance au DDT (Daborn et al., 2002). Cette surexpression est en réalité due à l'insertion d'un rétrotransposon, contenant des séquences régulatrices, en amont du gène (Chung et al., 2007) ; c'est donc une mutation purement régulatrice, dont la conséquence évolutive est directement observable.

A l'échelle génomique, plusieurs études inter-spécifiques ont analysé l'évolution des sites de liaison d'un facteur de transcription donné au cours de l'évolution (Dowell, 2010). Chez cinq espèces de vertébrés, les sites de liaison du facteur CEBPA ont été identifiés par ChIP-Seq (Schmidt et al., 2010), expérience qui consiste à immuno-précipiter les complexes facteur de transcription/ADN *in vivo* (Park, 2009). Chez des espèces assez proches comme l'homme et la souris, séparés par 80 millions d'années d'évolution, seulement 20% des sites de liaison de CEBPA sont en commun (Schmidt et al., 2010). L'analyse des sites de liaison des facteurs STE12 et TEC1 par ChIP-chip, chez trois espèces très proches de levure *Saccharomyces,* a révélé qu'à nouveau environ 20% des sites de liaison étaient conservés entre les trois espèces (Borneman et al., 2007). Ceci souligne la grande flexibilité des sites de

liaison d'un facteur de transcription au cours de l'évolution, même chez des espèces proches. Il faut néanmoins préciser que les éléments *cis* mutés dans cette dernière étude ne semblaient pas modifier significativement l'expression des gènes associés ; ces fréquentes mutations représentent peut-être d'avantage un réservoir de variabilité génétique que des véritables changements de régulation (Borneman et al., 2007).

c) Evolution des corégulateurs

Une autre possibilité théorique d'évolution du réseau, moins souvent mentionnée, est la modification des corégulateurs du facteur *trans* du réseau. Ce sont des protéines qui vont s'associer au facteur de transcription, et moduler sa fonction (Mannervik et al., 1999). Parfois, l'activité du facteur de transcription est impossible sans la présence de son corégulateur (si par exemple il lui apporte une fonction d'activation transcriptionnelle, comme c'est le cas pour LEAFY) (Parcy et al., 1998) ; d'autres fois, le facteur de transcription aura une activité différente selon le corégulateur présent. Ainsi, comme nous le verrons dans la deuxième partie de l'introduction, *LEAFY* est exprimé dans l'ensemble du méristème floral (la structure qui initiera les futures fleurs), et est pourtant capable de spécifier l'identité de chacun des organes floraux dans des territoires bien déterminés, en fonction du corégulateur qui y est exprimé (Lee et al., 1997; Lohmann et al., 2001; Liu et al., 2009b). On peut facilement imaginer comment l'évolution a pu « jouer » sur le domaine d'expression de ces corégulateurs ou sur leurs propriétés pour conditionner l'activité régulatrice d'un facteur de transcription.

3) Prédire une régulation transcriptionnelle et son évolution

Pour comprendre l'évolution d'un processus contrôlé par un réseau de régulation connu chez une espèce modèle, il nous faudrait savoir quand ce réseau est apparu, si sa structure a évolué, et donc chez quelles espèces actuelles il est présent. Nous disposons de plus en plus de données génomiques chez de nombreuses espèces non modèles (plus de 2000 génomes séquencés entièrement d'après la Genome Online Database), mais en tirer des informations fonctionnelles reste très difficile. Est-on capable de prédire l'existence d'un réseau de régulation sur la base d'une information génomique ? Peut-on tout d'abord prédire la position des éléments *cis* reconnus par le facteur de transcription qui contrôle le réseau ?

a) Prédire la liaison d'un facteur de transcription à l'ADN

Pour pouvoir prédire le positionnement d'un facteur de transcription sur l'ADN, il faut disposer de sa spécificité de liaison à l'ADN, c'est-à-dire des bases nucléotidiques qu'il lie préférentiellement à chaque position de son site de liaison. Pour cela, différentes méthodes expérimentales peuvent être employées, dont voici les plus couramment utilisées :

- Le SELEX (Systematic Evolution of Ligands by EXponential enrichment) consiste à utiliser une banque d'oligonucléotides de séquence aléatoire, et à sélectionner, au cours de plusieurs cycles d'enrichissement, les oligonucléotides reconnus par la protéine d'intérêt (Tuerk and Gold, 1990; Djordjevic, 2007). J'ai utilisé cette méthode au cours de ma thèse ; je la détaillerai dans la partie résultats, en soulignant ses avantages et ses limitations.
- Les PBM (Protein Binding Microarrays) constituent une autre approche *in vitro,* utilisant une puce contenant toutes les séquences possibles d'un oligonucléotide d'une longueur donnée. La liaison de la protéine à un oligonucléotide est repérée par immunodétection (Mukherjee et al., 2004). Cette technique est rapide et peut théoriquement être employée pour n'importe quel facteur de transcription, mais les motifs représentés sont pour le moment limités à une longueur de 11 pb (Godoy et al., 2011), ce qui pourrait se révéler trop court pour certains facteurs de transcription.
- Les techniques de type ChIP (Chromatin Immuno Precipitation), suivies d'une hybridation sur microarray (ChIP-chip) ou d'un séquençage massif (ChIP-Seq) (Park, 2009) permettent de positionner les sites de liaison du facteur de transcription à l'ADN *in vivo*, et sont donc beaucoup plus informatives que les deux techniques *in vitro* citées précédemment. L'approche expérimentale est néanmoins assez lourde et nécessite de disposer d'un très bon anticorps dirigé contre la protéine d'intérêt.

La spécificité de liaison du facteur de transcription, suffisamment précise, peut être ensuite utilisée pour aller prédire ses sites de liaison dans une séquence génomique. Pour cela, une approche encore trop fréquemment utilisée dans la littérature consiste à rechercher uniquement la séquence « préférée » du facteur de transcription (qu'on appelle la séquence consensus) ; c'est par exemple celle qui a été isolée le plus grand nombre de fois en SELEX ou en PBM. Il est bien connu maintenant qu'un facteur de transcription est loin de reconnaître une séquence unique, mais peut se lier à tout un ensemble de séquences proches. Utiliser une séquence consensus pour prédire des sites de liaison est donc une approche extrêmement limitante, qui engendrera de nombreux faux négatifs (Schneider, 2002).

L'approche choisie dans cette étude utilise une matrice de fréquences, qui va représenter, pour chaque position du site de liaison, la fréquence de trouver chacun des quatre nucléotides possibles. Cette matrice pourra directement prédire l'affinité du facteur de transcription pour n'importe quelle séquence d'ADN (ceci sera détaillé dans le chapitre II des résultats). L'utilisation d'une matrice permet de dépasser l'approche binaire de l'utilisation du consensus, et de considérer tous les sites de liaison possibles du facteur de transcription.

b) De l'*in vitro* à l'*in vivo*

La liaison d'un facteur de transcription à l'ADN *in vivo* ne dépend pas uniquement de sa spécificité de liaison. Le positionnement des nucléosomes, qui peuvent entrer en compétition avec les facteurs de transcription (Segal and Widom, 2009), ou encore l'état de compaction de la chromatine (John et al., 2011) vont conditionner l'accessibilité de la protéine à ses sites de liaison. Ces données peuvent être intégrées aux modèles précédents, pour obtenir des prédictions plus performantes. La performance d'un modèle peut être évaluée par sa sensibilité (prédiction d'un grand nombre de sites) et sa spécificité (détection des vrais positifs). Ces valeurs atteignent respectivement 35,1% et 21,3% pour la prédiction des sites de liaison d'un ensemble de facteurs de transcription humains présents chez des lymphoblastes, par comparaison à leur liaison réelle en ChIP-Seq. En y intégrant des données de positionnement des nucléosomes et d'hypersensibilité à la DNase (qui révèle l'accessibilité à l'ADN), ces nombres peuvent atteindre respectivement des valeurs de 60,5% et 81,5%, accroissant donc très fortement le pouvoir prédictif du modèle (Pique-Regi et al.).

c) De la liaison à la régulation

Prédire le site de liaison d'un facteur de transcription est possible, mais comment savoir si ce site est fonctionnel ? Le saut entre liaison et régulation est encore très grand, et mal compris pour le moment. Il est néanmoins possible d'insérer dans le modèle des données de régulation génétique obtenues par micro-array (Bar-Joseph et al., 2003), ce qui améliore la prédiction des sites régulateurs.

Une autre approche assez classiquement utilisée, le phylogenetic footprinting (Gumucio et al., 1992), pourrait permettre de détecter des sites fonctionnels, selon le principe que de tels sites ont de fortes chances d'être conservés chez des espèces relativement proches. En alignant ainsi les régions promotrices de plusieurs gènes orthologues, on peut détecter des zones bien conservées que l'on pense correspondre à des sites régulateurs. Cette approche a

bien sûr ses limitations, et le nombre et le choix des espèces utilisées est crucial, puisqu'il faut disposer d'espèces suffisamment éloignées pour détecter la différence entre les régions à faible et à fort taux de mutation, mais suffisamment proches pour pouvoir aligner les régions régulatrices et repérer des zones d'homologie.

Nous avons vu que l'évolution des réseaux de régulation est un moteur de l'évolution, et que de nombreux efforts ont été développés pour comprendre cette évolution et pour la prédire. Lors de ma thèse, je me suis intéressée à l'évolution du réseau transcriptionnel contrôlé par le facteur de transcription LEAFY (LFY). Pourquoi s'intéresser à cette protéine : car elle a une fonction unique chez les plantes à fleurs, mais l'origine de cette fonction chez les plantes terrestres est totalement inconnue. La seconde partie de mon introduction va donc s'attacher à présenter le rôle de LFY chez les angiospermes.

II) LFY et le réseau floral chez les angiospermes

1) Présentation du facteur de transcription LFY chez *Arabidopsis thaliana*

a) LFY *in vivo*

Lors du développement végétatif d'une angiosperme, un ensemble de cellules indifférenciées situées à l'extrémité de la tige (apex) vont permettre son développement et celui des feuilles latérales : c'est le méristème apical caulinaire (ou shoot apical meristem, SAM). Lors de la transition florale, qui est induite par différents signaux extérieurs (photopériode, température) et endogènes, le SAM est converti en un méristème d'inflorescence, qui va développer sur ses côtés des méristèmes floraux. Chez *A. thaliana*, l'initiation de ces méristèmes est répétitive et théoriquement infinie ; on qualifie pour cela le méristème d'inflorescence d'indéterminé. Les méristèmes floraux sont quant à eux déterminés et vont initier les verticilles des différents organes floraux.

LFY a originellement été découvert chez le muflier (*Antirrhinum majus*), où le gène a été baptisé *FLORICAULA* (*FLO*) (Schwarz-Sommer et al.; Coen et al., 1990). *LFY* a rapidement ensuite été identifié chez *A. thaliana* (Schultz and Haughn, 1991; Weigel et al., 1992). Les mutants *flo* et *lfy* montrent, à la position attendue des fleurs, des structures ressemblant à des tiges, qui se développent de manière indéterminée. Cette indétermination montre que ces tiges ont des caractères d'inflorescences, signifiant que *LFY* est nécessaire à la

transition entre méristème d'inflorescence et méristème floral. (**Fig. 2-A1**) (Schultz and Haughn, 1991; Weigel et al., 1992). Au fur et à mesure du développement de la plante, ces tiges-inflorescences ressemblent de plus en plus à des fleurs, portant des organes proches de sépales et éventuellement d'un pistil, mais ces « fleurs » restent pourtant toujours stériles (Schultz and Haughn, 1991; Weigel et al., 1992).

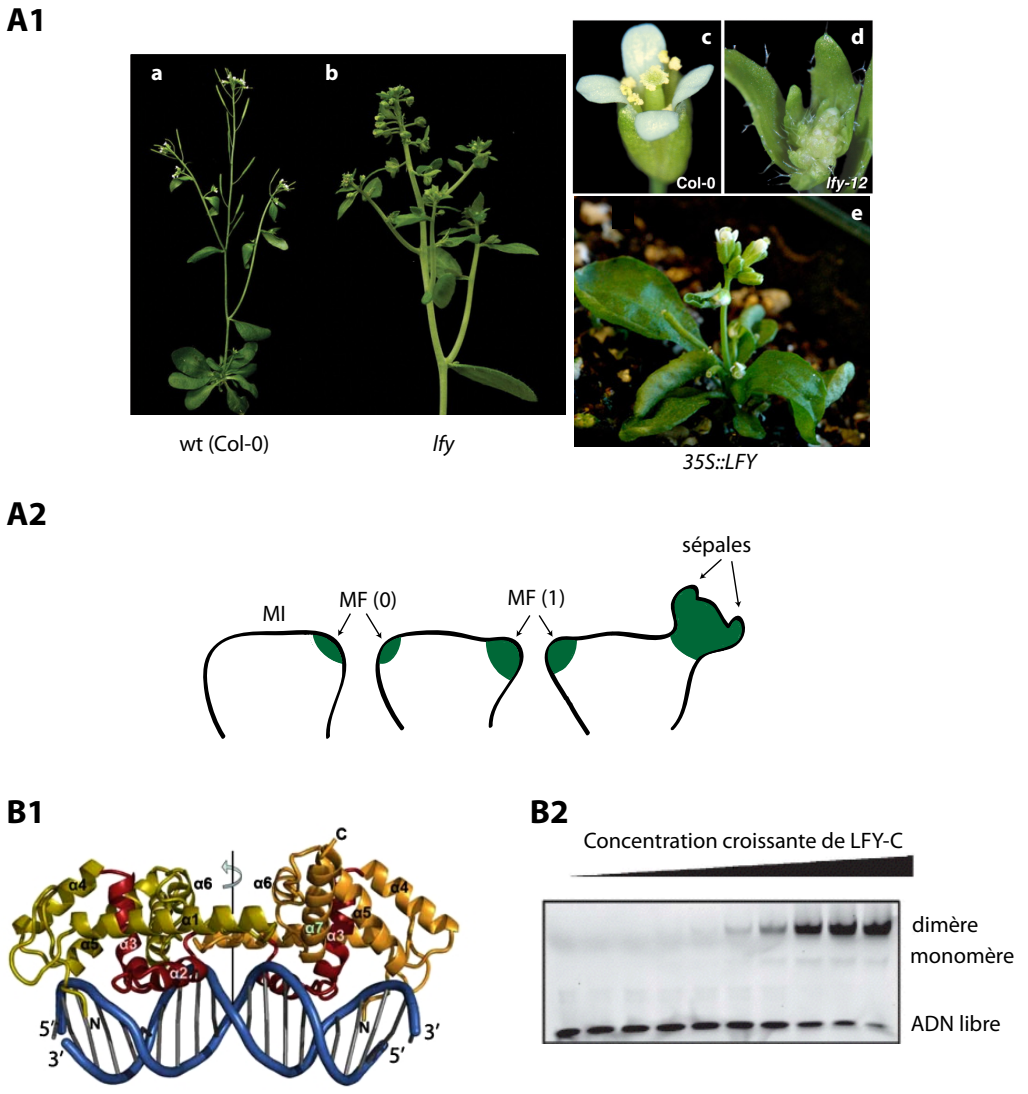


Figure 2 : Ensemble de données in *vivo* et *in vitro* présentant le rôle et les propriétés de LFY chez *A. thaliana*. A: Données fonctionnelles concernant *LFY in planta*. **(1)** Phénotypes des plantes sauvage (wt), mutante *lfy* et surexprimant *LFY* (*35S::LFY*). Vue d'ensemble de l'inflorescence (a, b, e), et détail de l'apex (c, d) pour les plantes wt et *lfy*. Origine des photographies : (Wu et al., 2003; Benlloch et al., 2007; Chae et al., 2008). **(2)** Patron d'expression de *LFY* (en vert) dans le méristème d'inflorescence (MI) chez *A. thaliana*. *LFY* est exprimé dans le méristème floral dès le stade 0 (MF (0)), puis dans la fleur jusqu'à l'émergence des différents organes floraux. Les stades sont définis d'après (Smyth et al., 1990). **B:** Données biochimiques. **(1)** Structure du domaine C-terminal de LFY (LFY-C, domaine de liaison à l'ADN), révélant sa dimérisation pour la liaison à l'oligonucléotide *AP1*. Un des monomères est colorié en jaune, et l'autre en orange, avec le domaine HTH colorié en rouge (Hamès et al., 2008). **(2)** Exemple de gel retard réalisé avec LFY-C sur l'oligonucléotide *AP1* (marqué grâce à un fluorophore): la bande correspondant à la liaison d'un monomère est bien plus faible que celle du dimère, témoignant de la coopérativité de liaison de LFY-C sur l'ADN (Hamès et al., 2008).

La fonction de *LFY in vivo* a progressivement été précisée par de nombreuses études. Ainsi, *LFY* et le gène MADS *APETALA1* (*AP1*) spécifient l'identité florale d'un méristème, et sont des antagonistes du gène *TERMINAL FLOWER 1* (*TFL1*), qui spécifie l'identité végétative et inflorescentielle (Shannon and Meek-Wagner, 1993). Les plantes surexprimant *LFY* ou *AP1* (*35S::LFY* ou *35S ::AP1*) montrent un phénotype de floraison très précoce (**Fig. 2-A1**) (Mandel and Yanofsky, 1995; Weigel and Nilsson, 1995; Liljegren et al., 1999), indiquant que *LFY* et *AP1* sont chacun suffisants pour induire la floraison et se placent à un niveau équivalent dans la voie de signalisation associée.

Le patron d'expression de *LFY* est cohérent avec sa fonction : pour des plantes cultivées en jours courts, pour lesquelles la transition florale se fait assez tardivement, on observe une augmentation progressive du niveau d'expression de *LFY* de l'état végétatif (dans les jeunes feuilles) jusqu'à l'état reproductif. Il existe donc un seuil d'expression de *LFY* qui va déclencher la transition florale (Blazquez et al., 1997). Après cette transition, *LFY* sera exprimé très tôt dans le méristème floral, puis dans les ébauches de chacun des organes floraux tour à tour (**Fig. 2-A2**) (Weigel et al., 1992; Blazquez et al., 1997).

b) LFY *in vitro*

Après les nombreuses études *in planta* qui ont précisé le rôle de LFY dans l'initiation de la floraison, diverses études biochimiques ont cherché à caractériser sa fonction moléculaire. Le fait que LFY contrôle *in vivo* l'expression de plusieurs gènes MADS impliqués dans le développement de la fleur (comme *APETALA1-AP1* et *AGAMOUS-AG*) suggérait déjà que LFY était un facteur de transcription, ce qui a été confirmé par la suite (Parcy et al., 1998).

LFY ne ressemble à aucun autre facteur de transcription connu, ce qui en fait une protéine unique dans le règne végétal (Maizel et al., 2005). En 2008, des membres de l'équipe ont résolu la structure cristallographique du complexe entre le domaine C-terminal de LFY (LFY-C, domaine de liaison à l'ADN) et un oligonucléotide *AP1* de 19 pb, constitué du site de liaison de LFY identifié chez *AP1* (**Fig. 2-B1**) (Hamès et al., 2008). Dans cette structure, LFY-C contacte l'ADN sous forme de dimère, l'oligonucléotide *AP1* étant constitué de deux demi-sites, chacun contacté par un monomère. La structure a révélé un repliement original de la protéine à 7 hélices-α, qui présente des similarités avec des protéines de type HTH (helix-turn-helix) telles que la protéine à homéodomaine Engrailed, suggérant que LFY a pu avoir une origine commune avec ces protéines. LFY-C contacte à la fois le squelette phosphate de

l'ADN et des bases nucléotidiques, assurant ainsi une spécificité de liaison sur un oligonucléotide de grande taille. Par des expériences de gel retard ou EMSA (Electrophoretic Mobility Shift Assay), qui consistent à faire migrer le mélange ADN-protéine sur un gel natif permettant la détection des complexes par retard de migration, Hamès et al. ont également pu montrer que la liaison du dimère de LFY-C sur l'ADN était coopérative (**Fig. 2-B2**). Les données structurales ainsi que l'alignement de 3 sites de liaison de LFY présents dans les gènes *AP1* et *AG* ont permis de proposer un consensus pour sa spécificité de liaison, du type (T/A)NNNNCCANTG(T/G)NNNN(T/A) (Busch et al., 1999; Hamès et al., 2008). De nombreux indices sont donc déjà disponibles sur les propriétés moléculaires de LFY, en tant que facteur de transcription.

2) Le réseau floral contrôlé par LFY

*J'ai participé à l'écriture d'une revue sur LFY lors de ma thèse (**Article 1**, p. 25), résumant ses rôles connus chez les plantes terrestres. Pour la suite de l'introduction, je vais donc rappeler les points essentiels à la compréhension des résultats, mais le lecteur pourra trouver de nombreuses informations complémentaires dans la revue, insérée à la fin de l'introduction.*

L'expression de *LFY* est activée par des protéines répondant à un ensemble de stimuli environnementaux ou endogènes de la floraison (Liu et al., 2009a; Posé et al., 2012). LFY, au sein de ce réseau, se positionne comme un intégrateur floral, qui va recevoir les informations issues des différentes voies de signalisation, et déclencher la formation du méristème floral (Blazquez and Weigel, 2000).

Lors du développement floral, LFY joue un double rôle. Il réprime tout d'abord le gène *TFL1* qui spécifie l'identité d'un méristème d'inflorescence (Shannon and Meek-Wagner, 1993; Ratcliffe et al., 1999), et induit les gènes qui déterminent l'identité florale comme *AP1*, *CAL* ou *LMI1* (Parcy et al., 1998; William et al., 2004; Saddic et al., 2006). Ces différents gènes constituent un réseau doté de nombreux rétrocontrôles positifs et négatifs qui assurent une transition florale brutale et irréversible (**Fig. 3A**). Ainsi *TFL1* réprime les gènes floraux (*LFY*, *AP1*) qui le répriment, alors que ces derniers s'activent entre eux (Liljegren et al., 1999; Ferrandiz et al., 2000; Siriwardana and Lamb, 2012).

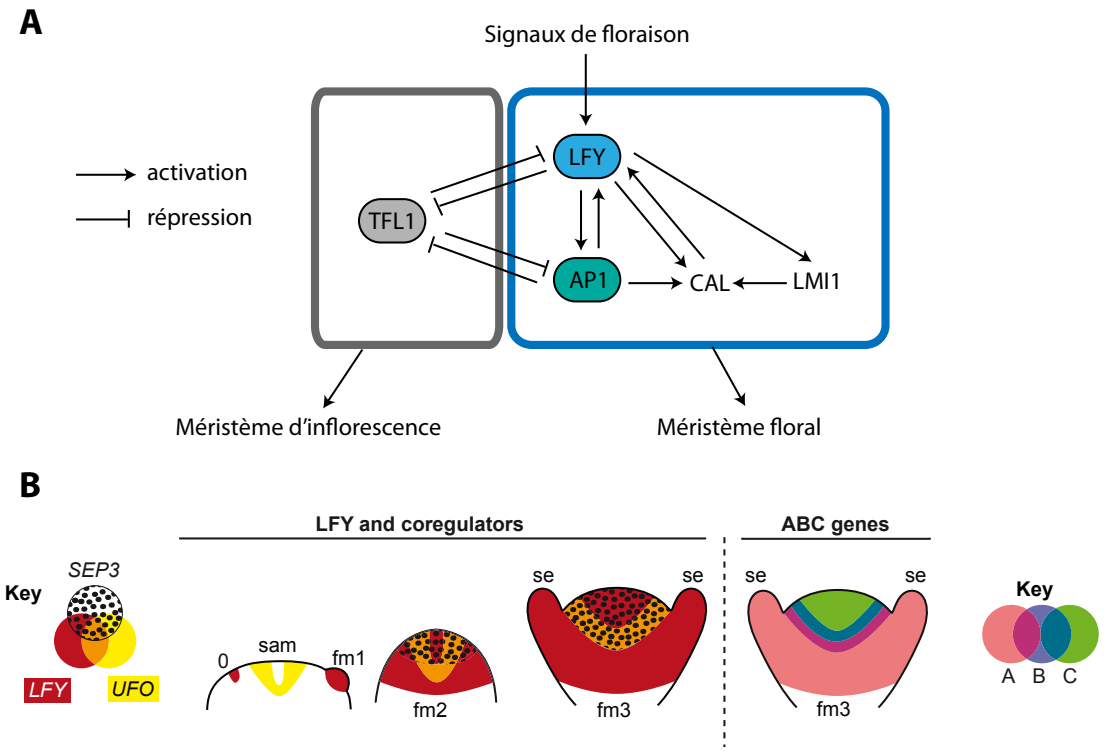


Figure 3 : Vue générale du réseau floral contrôlé par LFY. A : LFY participe en premier lieu à l'identité du méristème floral, en intégrant les signaux de floraison externes et internes. Il va ainsi activer ou réprimer un ensemble de gènes qui participent à la conversion irréversible du méristème d'inflorescence en méristème floral. **B** : Dans un deuxième temps, LFY active les gènes ABC, de manière séquentielle dans le temps et dans l'espace, grâce à l'expression différentielle de ses corégulateurs (SEP3, UFO) dans le méristème floral. La combinaison de l'expression des gènes ABC va spécifier l'identité des organes floraux. sam : shoot apical meristem (méristème apical caulinaire), fm : méristème floral, stades 0 à 3 comme définis dans (Smyth et al., 1990), se : sépales.

Par la suite, LFY intervient dans la compartimentation du méristème floral en induisant localement l'expression des gènes d'identité des organes floraux. Il active pour cela directement plusieurs gènes MADS : *AP1*, *APETALA3* (*AP3*), et *AGAMOUS (AG)* (Parcy et al., 1998; Busch et al., 1999). Ces facteurs de transcription, accompagnés des protéines APETALA2 (AP2) et PISTILLATA (PI), vont préciser l'identité des organes floraux dans des territoires bien déterminés ; ceci a conduit à proposer le modèle ABC pour expliquer les bases génétiques de cette identité (Coen and Meyerowitz, 1991). D'après ce modèle, l'expression, par exemple, des gènes A (*AP1*, *AP2*) seuls va déterminer un groupe de cellules à former un sépale, alors que l'expression combinée des gènes A et B (*AP3*, *PI*) va les déterminer à former un pétale (**Fig. 3B**). Ainsi, la combinatoire de l'expression spatiale des gènes A, B et C permet d'obtenir les quatre types d'organes floraux, dans des territoires bien précis. L'activation différentielle des gènes A, B et C par LFY dans des territoires différents, alors que *LFY* est exprimé dans l'ensemble du méristème floral, est possible grâce à la présence de différents corégulateurs dans ces territoires distincts (**Fig. 3B**). Ainsi, il a été proposé que

SEP3 et LFY interagissent pour activer les gènes B et C (Liu et al., 2009b), alors que les complexes LFY-UFO et LFY-WUS sont nécessaires à l'activation, respectivement, des gènes B ou C uniquement (Levin and Meyerowitz, 1995; Lee et al., 1997; Lohmann et al., 2001).

LFY est donc un facteur de transcription unique, au rôle clé dans le développement floral. LFY n'est pourtant pas uniquement présent chez les angiospermes, mais également chez d'autres groupes de plantes terrestres qui ne développent pas de fleurs. Son rôle dans le contrôle de la floraison est donc forcément dérivé : a-t-on des indices sur sa fonction, ses propriétés, ses gènes cibles avant l'apparition de la fleur ? Que connaît-on déjà de l'évolution du réseau transcriptionnel contrôlé par LFY chez les plantes terrestres?

III) Evolution de LFY et de son réseau chez les plantes terrestres

1) LFY chez les plantes « sans fleurs »

Les plantes terrestres sont classiquement séparées en quatre grands groupes : les mousses (ou bryophytes au sens large), les fougères (ou ptéridophytes au sens large), les gymnospermes et les angiospermes (**Fig. 4**). Bien que les mousses et les fougères ne soient pas des groupes monophylétiques (Qiu, 2008b) et que la monophylie des gymnospermes ne soit démontrée que depuis peu (Lee et al., 2011), ce regroupement est classiquement utilisé et s'est révélé pratique pour de nombreuses études.

LFY a été identifié chez des représentants de chacun de ces groupes. Chez l'ensemble des gymnospermes (à l'exception du genre *Gnetum*), il existe une seconde copie de *LFY*, nommée *NEEDLY* (*NLY*), qui a probablement été perdue lors de la radiation des angiospermes (Frohlich, 2003). C'est le seul cas de duplication de *LFY* qui ait perduré pendant une grande période (environ 150 millions d'années depuis la divergence angiospermes-gymnospermes) (Clarke et al., 2011). *LFY* et *NLY* sont exprimés dans des tissus végétatifs (SAM, feuilles ou aiguilles, tige, méristèmes axillaires,…) ainsi que dans les tissus reproductifs mâles ou femelles (Mellerowicz et al., 1998; Mouradov et al., 1998; Guo et al., 2005; Vazquez-Lobo et al., 2007; Shiokawa et al., 2008). Des gènes de type B et C sont parfois exprimés dans ces mêmes tissus reproductifs plus tardivement, suggérant que LFY/NLY pourraient réguler un réseau pré-floral de gènes MADS chez les gymnospermes (ceci sera discuté dans la partie II des résultats, avec une étude chez *Welwitschia mirabilis*). Les patrons d'expression de *LFY* et

NLY ne sont jamais identiques, ces deux gènes ont donc vraisemblablement des rôles distincts.

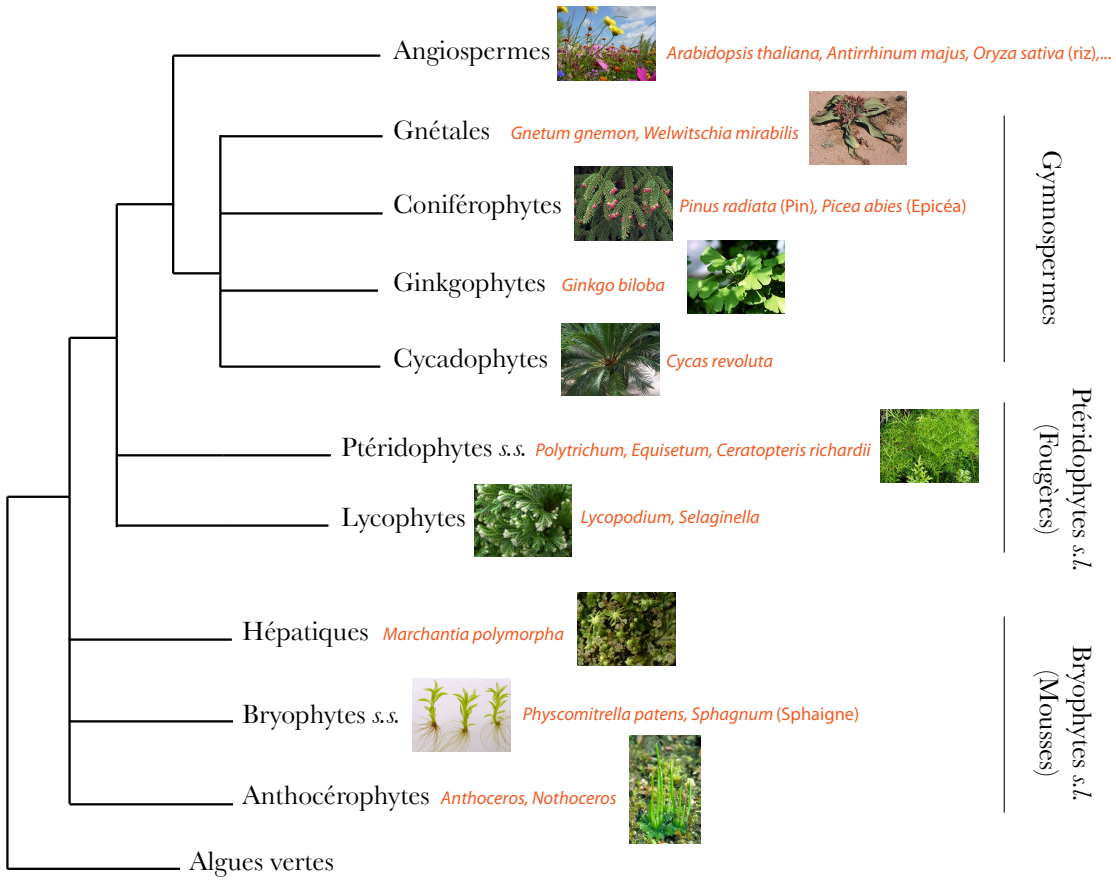


Figure 4 : Arbre phylogénétique simplifié des plantes terrestres, inspiré de (Frohlich and Chase, 2007). Les algues vertes sont choisies comme groupe externe. Les noms d'espèces représentatives de chaque groupe sont inscrits en rouge. *s.s.: sensus stricto* (sens strict), *s.l.: sensus lato* (sens large).

Chez les fougères, le domaine d'expression de *LFY* chez *Ceratopteris richardii* (*CrLFY1* et *CrLFY2*) a été examiné (Himi et al., 2001) : les deux copies sont exprimées dans des tissus végétatifs mais surtout reproductifs. Les patrons d'expression des gènes MADS étudiés ne coïncident pas avec ceux de *CrLFY1/2*, ce qui pourrait signifier que la régulation des gènes MADS par LFY est apparue après la divergence des fougères (Himi et al., 2001).

Chez les mousses, des études fonctionnelles ont pu être réalisées chez la plante modèle *Physcomitrella patens*. Cette espèce présente un cycle de développement en deux phases, l'une gamétophytique (haploïde) pendant laquelle se développe une tige feuillée d'où seront issus les gamètes nécessaires à la fécondation, puis une phase sporophytique (diploïde) qui produira les spores rétablissant l'haploïdie (Cove, 2005). *P. patens* possède deux copies de *LFY* (*PpLFY1* et *PpLFY2*), très proches en séquence protéique, et exprimées de manière similaire dans le gamétophyte (dans l'apex principal et les apex latéraux, et dans les structures reproductrices femelles) et dans l'ensemble du sporophyte (Tanahashi et al., 2005). Les deux

gènes présentent une forte redondance, et le double mutant *pplfy1 pplfy2* arrête son développement au stade zygote ; en effet, le développement gamétophytique et la fécondation se déroulent correctement, mais le zygote n'effectue pas sa première division cellulaire (Tanahashi et al., 2005). *PpLFY1*, exprimé sous le contrôle du promoteur de *LFY*, est incapable de complémenter un mutant *lfy* d'*A. thaliana*, à l'opposé de presque tous les autres gènes *LFY* issus d'angiospermes, de gymnospermes, de fougères ou même de certaines autres mousses (Maizel et al., 2005). Cet ensemble de résultats souligne donc une différence à la fois fonctionnelle et biochimique entre LFY et PpLFY1/2.

2) Un rôle ancestral pour LFY ?

Peut-on unifier toutes les fonctions de LFY présumées chez les plantes terrestres ? Chez les angiospermes, les orthologues de *LFY,* en plus de leur fonction florale, ont parfois un rôle « non floral » : chez le pois, *UNIFOLIATA* contrôle le caractère composé de la feuille (Hofer et al., 1997) et chez le riz, *RFL* détermine le degré de branchement de l'inflorescence et la croissance des talles (tiges secondaires spécifiques des céréales) (Kyozuka et al., 1998; Rao et al., 2008). Très souvent, le patron d'expression des orthologues de *LFY* suggère une fonction non reliée à la floraison : chez la vigne, *VFL* est exprimé dans tous les méristèmes, floraux ou végétatifs (Carmona et al., 2002) ; chez le peuplier, *PtLFY* est fortement exprimé dans la racine (An et al., 2011),...

Cette diversité, ainsi que les indices récoltés chez les gymnospermes, les fougères et les mousses, nous a amené à proposer un rôle unificateur pour *LFY*. *LFY* pourrait ainsi contrôler ancestralement la division cellulaire, l'extension cellulaire ou l'organisation des cellules en structures indifférenciées de type méristèmes, pendant la phase sporophytique (Moyroud et al., 2010). Ce rôle est directement observé chez *Physcomitrella patens*, et est supposé être semblable chez les fougères. Chez les gymnospermes et les angiospermes, ce rôle ancestral aurait été restreint à des régions méristématiques (méristèmes axillaires, marge de la feuille composée,…), zones actives de division et de réarrangement cellulaire. *LFY* aurait en parallèle acquis le contrôle d'un réseau reproductif (réseau floral chez les angiospermes), par le biais de la régulation des gènes MADS. Ce scénario reste hypothétique mais permettrait d'intégrer les données très variées observées chez les plantes terrestres. Le réseau « dérivé » contrôlé par LFY serait donc assez bien connu (au moins chez les angiospermes), mais son réseau « ancestral », s'il existe, resterait indéterminé.

3) Evolution de LFY et de son réseau

Des données sur l'évolution de LFY, en tant que facteur *trans* du réseau, peuvent nous permettre d'appréhender l'évolution du réseau qu'il contrôle. En effet, alors que les gènes régulateurs évoluent classiquement par duplication, *LFY* est en copie unique chez la majorité des plantes terrestres, et certains des cas où plusieurs copies de *LFY* existent sont reliés à des évènements de polyploïdisation du génome **(Fig. 5)**, comme par exemple chez les céréales (Adams and Wendel, 2005). *LFY* n'a néanmoins jamais formé de famille multigénique semblable à celle formée par les gènes MADS, et tend à être conservé à l'état d'une ou deux copies chez l'ensemble des plantes terrestres.

Division	Classe ou sous-division	Espèce	Nombre de gènes orthologues *LFY*	Etat de ploïdie du génome	Nom du (des) orthologue(s) de *LFY*
Angiospermes	Eudicotylédones	*Arabidopsis thaliana*	**1**	**2n**	*LEAFY (LFY)*
		Antirrhinum majus	**1**	**2n**	*FLORICAULA (FLO)*
		Nicotiana tabacum	**2**	**4n**	*NFL1, NFL2*
		Petunia x hybrida	**1**	**2n**	*ALF*
		Malus domestica	**3 ou plus**	**polyploïde***	*AFL1, AFL2, ?*
		Mimulus guttatus	**2**	**2n**	*MimGuFLOA, MimGuFLOB*
		Verbena officinalis	**2**	**2n**	*VerOfFLOA, VerOfFLOB*
		Vitis vinifera	**1**	**2n**	*VFL*
		Pisum sativum	**1**	**2n**	*UNIFOLIATA (UNI)*
	Monocotylédones	*Oryza sativa*	**1**	**2n**	*RFL*
		Zea mays **	**2**	**4n**	*ZFL1, ZFL2*
	clade ANITA	*Amborella trichopoda*	**1**	**2n**	*AmboLFY*
		Nymphaea odorata	**1**	**6n**	
		Peperomia sp.	**1**	***	*PepspLFY*
Gymnospermes ****	Gnétales	*Welwitschia mirabilis*	**2**	**2n**	*WelLFY, WelNDLY*
		Gnetum gnemon	**1**	**2n**	
	Conifères	*Pinus radiata*	**2**	**2n**	*PRFLL, PrNDLY*
		Picea abies	**2**	**2n**	*PaLFY, PaNLY*
	Ginkgophytes	*Ginkgo biloba*	**2**	**2n**	*GinLFY, GinNdly*
Fougères (Ptéridophytes *s.l.*)	Ptéridophytes *s.s.*	*Ceratopteris richardii*	**2**	**2n**	*CrLFY1, CrLFY2*
		Psilotum nudum	**1**	**2n**	*PFY*
		Equisetum arvense	**2**	**2n**	*EaLFY1*
	Lycophytes	*Isoetes asiatica*	**1**	**2n**	*IsoLFY*
Mousses (Byrophytes *s.l.*)	Hépatiques	*Marchantia polymorpha*	**1**	**2n**	*MarpoFLO*
	Bryophytes *s.s.*	*Physcomitrella patens*	**2**	**2n**	*PpLFY1, PpLFY2*
		Atrichum angustatum	**2**	**2n**	*AtranFLO1, AtranFLO2*

Figure 5 : Nombre de copies du gène *LFY* chez différentes espèces des grands groupes de plantes terrestres, en relation avec l'état de ploïdie du génome. *Le pommier est un polyploïde complexe, son génome étant issu de nombreux évènements de duplications entières et partielles du génome. **Une duplication récente du génome a eu lieu chez le maïs. ***L'état de ploïdie du génome de *Peperomia* dépend de l'espèce étudiée. ****Chez toutes les gymnospermes (à l'exception du genre *Gnetum*), il existe un paralogue de *LFY* appelé *NEEDLY* (*NLY*).

La protéine LFY est également très fortement conservée en séquence chez les plantes terrestres (**Fig. 6**). Un alignement de nombreuses séquences entières de LFY permet d'identifier deux domaines : le domaine C-terminal, impliqué dans la liaison à l'ADN (Maizel et al., 2005; Hamès et al., 2008) et très conservé ; et le domaine N-terminal, au rôle encore

inconnu, assez bien conservé. Les 14 acides aminés impliqués dans le contact ADN-protéine, identifiés grâce à la structure de LFY-C lié à l'oligonucléotide *AP1* (Hamès et al., 2008), sont parfaitement conservés chez toutes les plantes terrestres (**Fig. 6**). Cette conservation, associée au fait que *LFY* n'a pas formé de famille multigénique, suggère que le réseau contrôlé par LFY n'a pas évolué en *trans*.

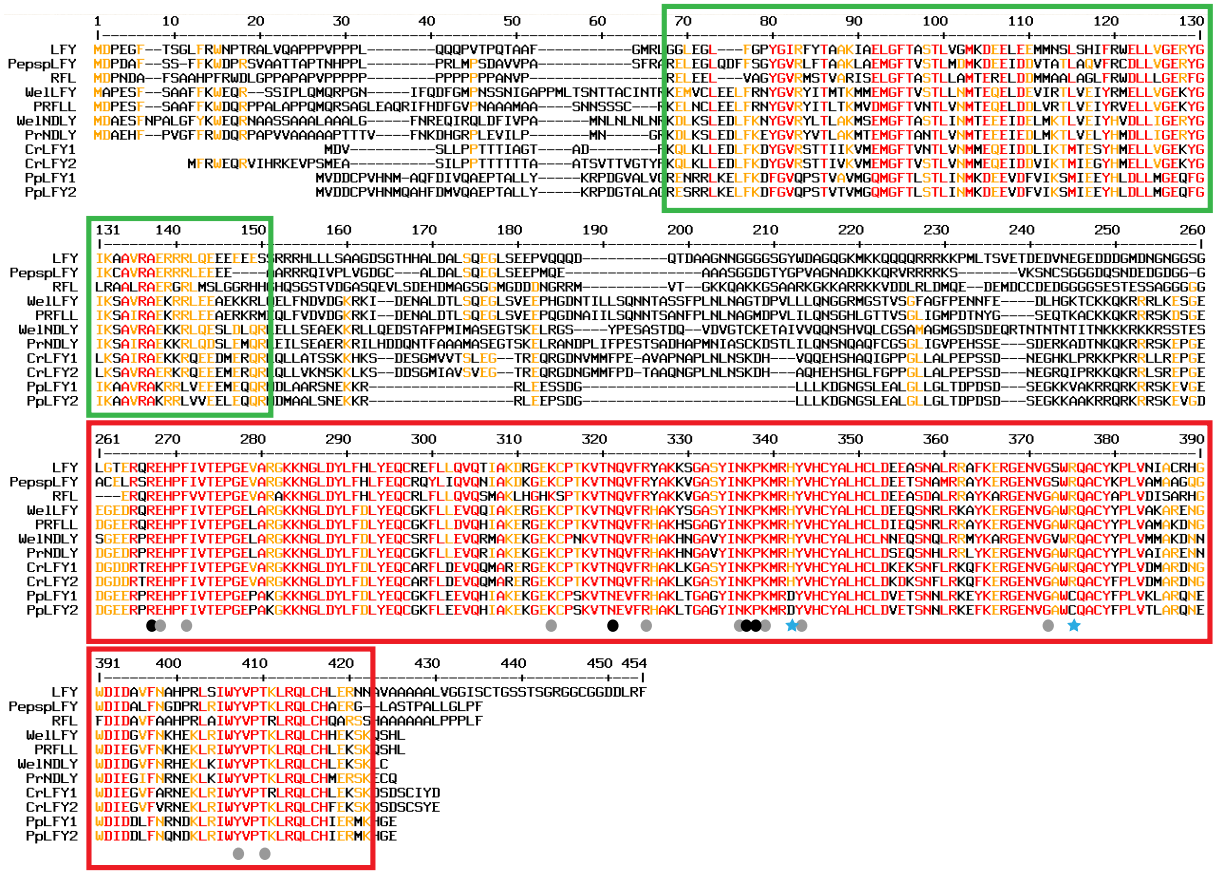


Figure 6 : Alignement des séquences protéiques entières de LFY chez différentes espèces de plantes terrestres. Les séquences choisies sont LFY, RFL et PepspLFY pour les angiospermes, WelLFY, WelNDLY, PRFLL, PrNdly pour les gymnospermes, CrLFY1 et CrLFY2 pour les fougères, et PpLFY1 et PpLFY2 pour les mousses. L'alignement est réalisé grâce au logiciel MultiAlin (http://multalin.toulouse.inra.fr/multalin/), les positions coloriées en rouge et en orange sont conservées respectivement à 90% et à 50% chez toutes les espèces choisies. Le domaine N-terminal est encadré en vert, et le domaine C-terminal, très conservé, est encadré en rouge. Les acides aminés impliqués dans le contact de LFY d'*A. thaliana* avec les bases nucléotidiques sont indiqués par un rond noir, et ceux impliqués dans le contact avec le squelette phosphate de l'ADN sont indiqués par un rond gris (Hamès et al., 2008). Les résidus divergents chez PpLFY1 et PpLFY2 (Maizel et al., 2005), et dont la fonction sera approfondie plus tard dans cette thèse, sont indiqués par une étoile bleue.

Pourtant, PpLFY1/2 ne reconnaît pas les séquences liées par LFY d'*A. thaliana* (Maizel et al., 2005). En outre, ce changement a une grande importance fonctionnelle puisque nous avons vu précédemment que PpLFY1 ne pouvait pas complémenter un mutant *lfy,* alors que la mutation d'un acide aminé de PpLFY1, lui permettant de reconnaître les séquences cibles de LFY, rend la protéine capable de complémenter le mutant (Maizel et al., 2005). Il

existe donc au moins un changement en *trans* dans l'histoire de l'évolution de LFY, ce qui était pourtant indétectable sur la base de la conservation des 14 acides aminés cités précédemment. Existe-t-il des changements similaires chez d'autres plantes terrestres ?

Review

LEAFY blossoms

Edwige Moyroud[1], Elske Kusters[2], Marie Monniaux[1], Ronald Koes[2] and François Parcy[1]

[1] Laboratoire de Physiologie Cellulaire Végétale, UMR5168, Centre National de la Recherche Scientifique, Commissariat à l'Énergie Atomique, Institut National de la Recherche Agronomique, Université Joseph Fourier, 17 av. des Martyrs, bât. C2, 38054 Grenoble, France
[2] Department of Molecular Cell Biology, Graduate School of Experimental Plant Sciences, VU-University, de Boelelaan 1085, 1081HV Amsterdam, The Netherlands

The *LEAFY* (*LFY*) gene of *Arabidopsis* and its homologs in other angiosperms encode a unique plant-specific transcription factor that assigns the floral fate of meristems and plays a key role in the patterning of flowers, probably since the origin of flowering plants. *LFY*-like genes are also found in gymnosperms, ferns and mosses that do not produce flowers, but their role in these plants is poorly understood. Here, we review recent findings explaining how the LFY protein works and how it could have evolved throughout land plant history. We propose that LFY homologs have an ancestral role in regulating cell division and arrangement, and acquired novel functions in seed plants, such as activating reproductive gene networks.

LEAFY: a master regulator of flower development

Mutations, such as *floricaula* (*flo*) in snapdragon (*Antirrhinum majus*) and *leafy* (*lfy*) in *Arabidopsis*, allowed the identification of a unique type of transcription factor that specifies floral identity of meristems and controls the very first steps in the formation of a flower (Figure 1). Following the isolation of *FLO* and *LFY*, homologs have been identified from numerous species, including gymnosperms and non-seed plants. Here, we summarize the recent findings enlightening the key role of *LFY* genes in multiple species with a variety of different floral architectures, as well as novel data illustrating how this unique protein works.

In flower meristems, LFY acts as a master regulator orchestrating the whole floral network: it activates downstream genes that give their unique identities to the floral meristem and the floral organ primordia [1–6]. It also controls the expression of several additional genes of unknown function, some of which might specify other floral traits such as whorled phyllotaxis or absence of internode elongation [7,8].

LFY is a plant-specific transcription factor that directly binds to the regulatory region of its target genes through a helix–turn–helix motif buried within a unique protein fold [9]. Surprisingly, *LFY* is found as a single gene in most land plant species, even in monocots where several events of gene duplications occurred [4]. LFY is uniformly expressed in floral buds and, therefore, regional coregulators are needed to induce distinct target genes in specific subdomains (Figure 2). The MADS box gene *SEPALLATA3 (SEP3)*, which is expressed in the three central floral whorls, acts as such a coactivator for the induction of B (*APETALA3*, *AP3*) and C (*AGAMOUS*, *AG*) floral organ identity genes [10]. The expression of *AG* is restricted to the flower center by several additional coactivators and corepressors [3,6]. To activate *AP3* in whorls 2 and 3, LFY binds to an F-box protein, known as UFO (UNUSUAL FLORAL ORGANS) in *Arabidopsis*, which is part of an SCF (Skp1–cullin–F-box protein)-type ubiquitin ligase [11]. How SCF^{UFO} promotes LFY activity is unclear, but is likely to be similar to the activation of a variety of transcription factors in yeast by the ubiquitination–proteasome system [12]. Several recent findings indicate that UFO is involved in the activation of many other LFY targets: mutation of the *UFO* homolog *FIMBRIATA* (*FIM*) in snapdragon reduces expression of both B and C genes [13] and in petunia (*Petunia hybrida*), tomato (*Solanum lycopersicon*) and lotus (*Lotus japonicus*), it causes a nearly complete loss of floral meristem identity and strongly perturbs the expression of all floral organ identity genes [14–16]. In addition, gain-of-function experiments have shown that constitutive expression of both *ABERRANT LEAF AND FLOWER* (*ALF*, petunia *LFY*) and *DOUBLE TOP* (*DOT*, petunia *UFO*) ectopically activates a wide spectrum of B-, C-, D- and E-type organ identity genes in petunia [15]. The relatively weak phenotype of *Arabidopsis ufo* and snapdragon *fim* mutants is probably due to functional redundancy, because LFY activity is fully dependent on coexpression of *UFO* or *DOT* when expressed in petunia [15]. Moreover, expression of a dominant negative form of *UFO* in *Arabidopsis* causes a strong *lfy*-like phenotype, including loss of floral meristem identity, which is similar to petunia and tomato *ufo* mutants [11].

LFY is also involved in grass flower development. In maize (*Zea mays*), the *LFY* homologs *ZFL1* and *ZFL2* are required for proper expression of B and C genes in flowers [17], whereas in rice (*Oryza sativa*), the *LFY* ortholog *RFL* is not expressed in floral meristems (Figure 2), and flowers appear fertile even when *RFL* is silenced [18,19]. However, the phenotype of these flowers has not been precisely described, and, because the *UFO* ortholog *ABERRANT PANICLE ORGANIZATION 1* (*APO1*) controls flower development in rice [20,21], RFL and APO1 might perform this task together. The *LFY* gene could therefore have had its floral function before the divergence of dicots and monocots, but this role was partially lost in rice and was taken over by another factor (e.g. the MADS box factor MOSAIC FLORAL ORGANS1 [22]). In Californian poppy

Corresponding authors: Koes, R. (ronald.koes@falw.vu.nl); Parcy, F. (francois.parcy@cea.fr).

 1360-1385/$ – see front matter doi:10.1016/j.tplants.2010.03.007 Trends in Plant Science 15 (2010) 346–352

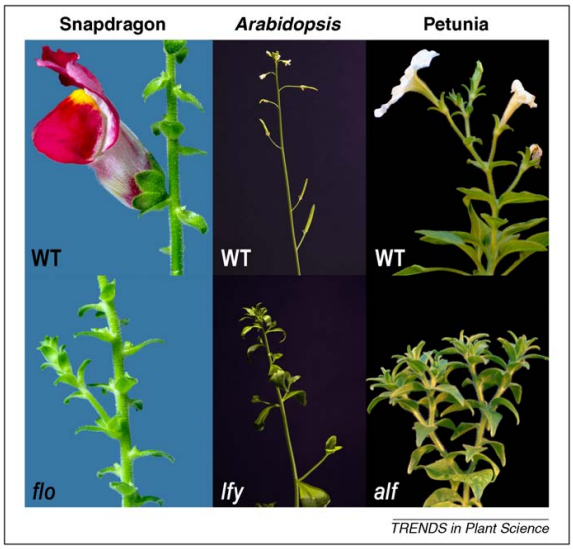


Figure 1. Phenotypes of wild type inflorescences in snapdragon, *Arabidopsis* and petunia and the corresponding *flo*, *lfy* and *alf* mutant.

(*Eschscholzia californica*), the expression pattern of the *LFY* ortholog *EcoFLO* does not coincide with expression of *AG* orthologs [23], suggesting that LFY might not regulate C genes in early branching eudicots. Studies in basal angiosperms will be crucial to understand whether LFY already controlled the floral network as a whole in the most recent common ancestor of flowering plants or whether ABC genes progressively came under LFY regulation as angiosperms diversified.

Role of *LFY* in the patterning and evolution of inflorescences

Angiosperm inflorescences consist of either solitary flowers or clusters of flowers with a variety of architectures [24,25] (Figure 3). In (open) racemes, the apical meristem grows indefinitely (i.e. it is indeterminate) and flowers develop from lateral meristems. Cymes show an opposite mode of development because the apical meristem terminates by forming a flower and growth continues from a lateral ('sympodial') meristem that generates the next unit. Panicles take an intermediate position since both apical and lateral meristems form flowers after several branching events. Given the importance of *LFY* in specifying floral meristem identity, it is not surprising that the spatiotemporal regulation of *LFY* activity is a major factor that determines when (flowering time) and where (inflorescence architecture) flowers are formed.

In *Arabidopsis*, which develops an open raceme, the time and the place where flowers form is indeed primarily regulated via the transcription of *LFY*. *LFY* is expressed during the vegetative phase in leaf primordia at steadily

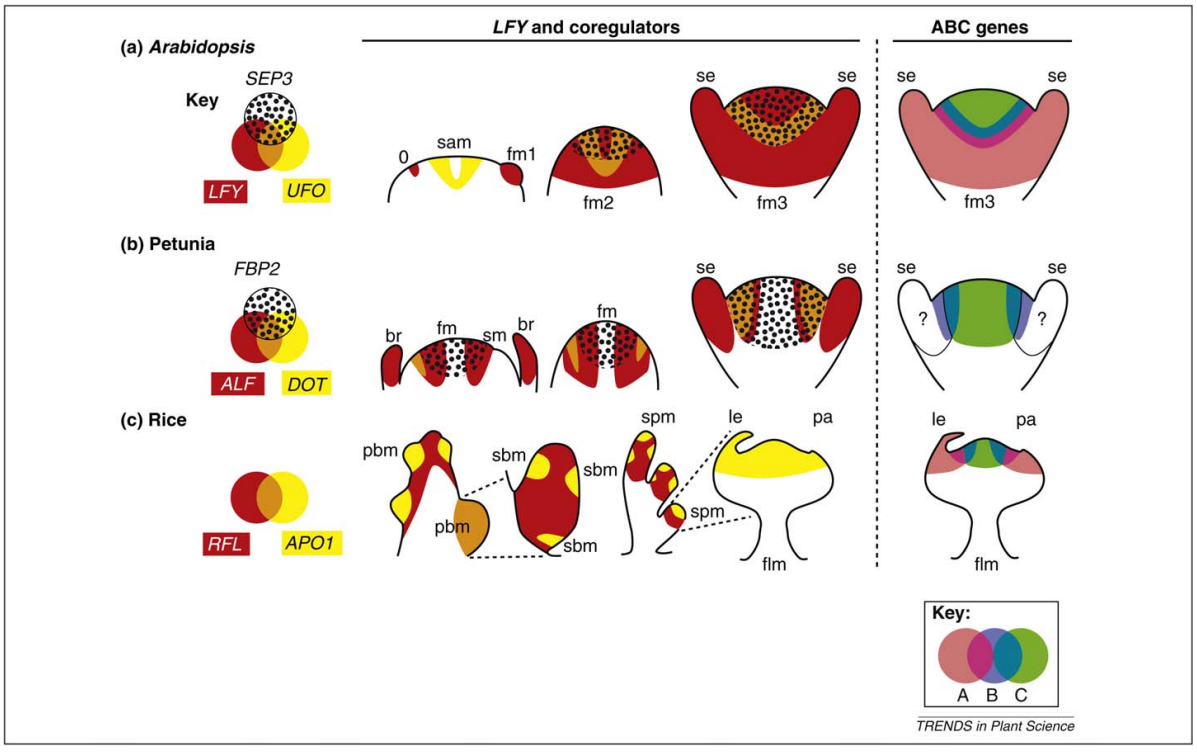


Figure 2. Expression patterns of *LEAFY* genes, some of their coregulators and their main target genes in **(a)** *Arabidopsis*, **(b)** petunia and **(c)** rice. Developing structures leading to early floral bud are schematically described. The color codes are explained in the keys. In rice, glumes have been omitted and only florets are shown. Flower meristems of stages 1, 2 and 3 (fm1, fm2 and fm3, respectively) are numbered according to Ref. [70]. Expression patterns for *LFY* and *UFO* genes have been depicted according to Refs [2,15,18,20,21,31,32]. The expression patterns of the ABC genes have been adapted from Refs [71–74]. Abbreviations: br, bract; flm, floret meristem; fm, floral meristem; le, lemma; pa, palea; pbm, primary branch meristem; sam, shoot apical meristem; sbm, secondary branch meristem; se, sepal; sm, sympodial meristem; spm, spikelet meristem.

Trends in Plant Science Vol.15 No.6

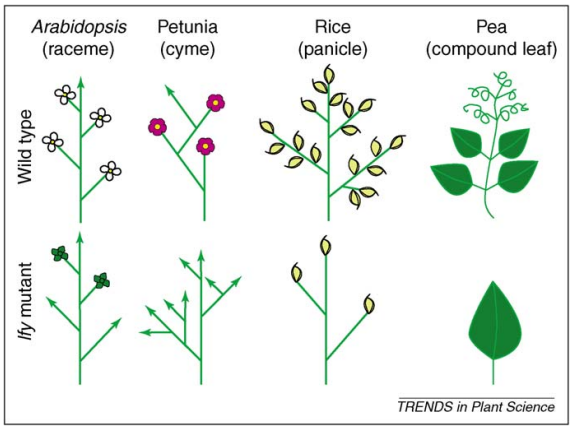


Figure 3. Inflorescence structures in *Arabidopsis*, petunia and rice and in pea leaf shape in wild type and *leafy* mutant plants. Green arrows indicate shoots topped with an indeterminate inflorescence meristem. Green flowers in *Arabidopsis* indicate a shoot or flower intermediate and abnormal flowers.

increasing levels until a threshold is reached, and flowering is triggered via the direct activation of *APETALA1* (*AP1*) [1,26]. Within the inflorescence, *LFY* is expressed in lateral (floral) meristems, whereas it is repressed in the apical inflorescence meristem by *TERMINAL FLOWER 1* (*TFL1*) [27]. Constitutive *LFY* expression converts both apical and axillary meristems into terminal flowers, indicating that the transcriptional regulation of *LFY* is the limiting factor defining when and where flowers are produced [28].

In cymes, floral identity is first specified in apical floral meristems, whereas it is transiently repressed in lateral sympodial meristems, which eventually also acquire floral identity after generating the next lateral primordium. Cymes are common among Solanaceae, and the *LFY* orthologs *NFL* from tobacco (*Nicotiana tabacum*), *FALSIFLORA* (*FA*) from tomato and *ALF* from petunia are expressed in different patterns than *LFY* and *FLO* in the racemose inflorescences of *Arabidopsis* and snapdragon. During early vegetative stages, *NFL*, *FA* and *ALF* are already highly expressed in (incipient) leaf primordia, whereas in the inflorescence meristems their mRNAs appear in the apical region and with a slight delay in the emerging lateral meristem that forms the next sympodial inflorescence unit [29–31]. However, these distinct solanaceous *LFY* transcription patterns do not account for the distinct cymose inflorescence architecture, because constitutive expression of either *LFY* or *ALF* does not alter the inflorescence or flowering time of petunia [15]. Here, LFY activity is restricted in time and space via the *UFO* homolog: *DOT* mRNA is only expressed during flowering and first appears in the apical (floral) meristem, whereas expression in the sympodial meristem is delayed, much more than that of *ALF* (Figure 2). Moreover, ectopic expression of either *DOT* or *UFO* is sufficient to trigger precocious flowering and convert the cymose inflorescence into a solitary flower [15]. In contrast, *Arabidopsis* meristems at the apex of embryos, vegetative shoots and inflorescences express *UFO*, but do not acquire floral identity, because the transcription of *LFY* is limiting, and, for the same reason, constitutive expression of *UFO* does not alter flowering time or inflorescence architecture in *Arabidopsis* [28,32]. Thus, the divergent spatiotemporal expression of floral meristem identity in these cymes and racemes seems to be largely the result of differences in both *LFY* and *UFO* homologs expression patterns (Figures 2 and 3) and is associated with a shift in the restriction of LFY activity in time and space between transcriptional and post-translational control. The directionality of this shift (back, forth or back and forth multiple times) cannot be inferred from the current data on petunia (a Rosid) and *Arabidopsis* (an Asterid) only and requires analysis of additional species across smaller phylogenetic distances.

It is thought that alterations in the expression patterns of developmental genes are a major factor driving the morphological divergence of organisms, but controversy exists as to whether this results mostly from changes in upstream regulatory proteins or, in another species, in the *cis*-regulatory DNA elements that control transcription [33–36]. Current data indicate that both types of alterations could have contributed to the divergence of *LFY* expression in angiosperms. Several species of the Brassicaceae, to which *Arabidopsis* belongs, express their *LFY* homologs within the apical meristem, which, nevertheless, remains indeterminate and does not convert into a flower [37,38]. Whether that is because these species do not express their *UFO* homologs in the apical meristem has not been examined. Studies in which promoter reporter gene constructs were introduced into *Arabidopsis* indicated that the expression pattern of the homologs of *LFY* in three other Brassicaceae are divergent from that of *LFY* in *Arabidopsis*, either because of an alteration in *cis*-regulatory elements (possibly a loss of the element responding to *TFL1*) or because of a yet unidentified difference in the upstream regulatory network [38,39]. The latter difference could involve any of the recently discovered regulators of *LFY* expression [40–44].

LEAFY function during grass inflorescence branching and legume leaf development

In rice, the *LFY* homolog, *RFL*, also controls inflorescence structure, but in a different way than in dicots. The rice inflorescence architecture (a panicle, Figure 3) is made of primary and secondary branches that hold the flowering structures (spikelets). Therefore, there are additional meristem identities in grasses that are absent in model eudicot inflorescences. *RFL* is expressed in the early panicle meristem and is required, together with *APO1*, to maintain its indeterminacy: downregulation of *RFL* or mutation of *APO1* reduce branch number and trigger precocious branch termination with spikelets [18–21]. *RFL* also controls branching at the whole plant level because silencing of *RFL* abolishes the development of tillers (secondary shoots growing from the base of the plant) [19]. As in rice, *ZFL* genes control the branching of the male inflorescence of maize (called the tassel) [17]. These findings show that *LFY* orthologs are able to promote meristematic (indeterminate) growth in grass inflorescences.

Interestingly, a related role was observed in a specific clade of legumes including pea (*Pisum sativum*), lotus

(*Lotus japonicus*) and alfalfa (*Medicago sativa*) [45]. These species generate compound leaves, each consisting of several leaflets, a process that involves the maintenance of a transient meristematic state at the margin of the developing leaves and in some species requires *LFY* activity. For example, when *UNIFOLIATA* (*UNI*, the pea *LFY* ortholog) is mutated [46], the leaves are simpler, sometimes with only one leaflet, showing that *UNI* is required, together with the *UFO* ortholog *Stamina pistilloida* [47], to promote the transient meristematic state required to form multiple leaflets (Figure 3). In other plants with compound leaves, leaf dissection is controlled by *KNOX* genes [45,48], which are also known to play an important role in apical meristem growth.

LFY genes, therefore, appear to promote an indeterminate and meristematic growth in both grass inflorescences and the leaves of some legumes. This could either represent the acquisition of a new function of *LFY* in angiosperms or an ancient role that exists (although sometimes hidden) in a wide array of plants species. Recent results in *Arabidopsis* support the latter hypothesis. Combining mutations in *PENNYWISE* (*PNY*) and *POUND–FOOLISH* (*PNF*) genes led to the production of cauline leaves devoid of axillary meristems. This phenotype is, at least in part, due to the strong *LFY* downregulation because constitutive *LFY* expression in such plants is able to direct flower development in the axils of leaves, as it does in the wild type [40]. Moreover, *pny pnf/+lfy* plants show a high proportion of cauline leaves lacking axillary meristem, a phenotype not seen in a *pny pnf/+* background [40]. These data nicely show that, in *Arabidopsis* too, *LFY* is able to stimulate meristem growth (in this case axillary meristems) and this function might be important during the early development of floral meristems before they are converted into flowers. We speculate that the role of *LFY* on the development of compound leaves, grasses inflorescence or *Arabidopsis* axillary meristems might reveal, in diverse manners, its capacity to stimulate meristematic growth. We thus propose that *LFY* possesses two functions: promoting meristem growth and conferring floral identity. Depending on the species, the two functions could be obvious (as in maize or pea), the first one might be cryptic (as in *Arabidopsis*) or the second one might be reduced (as in rice).

LEAFY in gymnosperms

In addition to the ortholog of *LFY*, the four groups of extant gymnosperms (ginkgo, cycads, gnetales and conifers) possess a related gene, *NEEDLY* (*NLY*) (first identified in a pine species, *Pinus radiata*, [49]), that was probably lost in the lineage leading to angiosperms [50]. The functions of the two proteins are unknown owing to the lack of gymnosperm mutants, but expression data exist that provide insights into the role of the family in non-flowering plants.

Gymnosperms do not form flowers but display their male and female organs on two distinct axes called strobili or cones, with the exception of some gnetalean taxa that have bisexual compound cones. *LFY* and/or *NLY* expression has been analyzed in six conifer species, one gnetale and *Ginkgo biloba* [49,51–57] and is commonly seen in vegetative tissues such as shoot apical meristem, stem and leaves or needles, but their role in these tissues is unknown. One interesting feature that is shared by all the gymnosperms that have been examined so far is the upregulation of *LFY* and/or *NLY* in axillary meristems, independently of their vegetative or reproductive fate. This suggests that the two proteins could be involved in the establishment of lateral meristems but are, in themselves, not sufficient to confer a reproductive status. *AGL6*-like genes, which form a clade that is sister to *SEP* genes [58] and have a role similar to *SEP3* in rice and petunia [22,59], are expressed in reproductive meristems of conifers and gnetales and, thus, are possible candidates to act together with LFY in the specification of reproductive identity in gymnosperms [51].

Do *LFY* genes also regulate MADS-box gene expression in gymnosperms? Whereas A genes and *UFO* orthologs have not been described in gymnosperms, orthologs of B and C genes do exist and are expressed in male and female gymnosperm structures, suggesting that they might, as in angiosperms, contribute to specify their identity [60]. Evidence that B and C orthologs are also regulated by LFY or NLY remains circumstantial. In early developmental stages of the reproductive meristem, expression of both *LFY* and *NLY* in nascent primordia generally precedes and encompasses that of B and C genes, consistent with a possible regulatory role. In later stages, *LFY* and *NLY* domains sometimes stop overlapping to become mutually exclusive. For example, in Norway spruce (*Picea abies*) male cones, *PaLFY* is detected in the sporogenous cells only, whereas *PaNLY* continues to be expressed in the surrounding tissues [57,61]. In female cones of at least three conifer species, *LFY* remains expressed in the ovule primordia coinciding with the C gene *DAL2*, whereas *NLY* is expressed in complementary tissues (cone axis and sterile scale) that are devoid of *DAL2* transcripts, suggesting that the two paralogs could regulate distinct sets of genes.

Gymnosperms with their two *LFY*-like genes that persisted for an extended period of time are an exception among land plants. This peculiarity could be explained by the distinct expression patterns of the two paralogs and possibly a different role. Indeed, *NLY* could have acquired a novel function and diverged slightly from LFY as suggested by the observation that *NLY* is not as efficient as gymnosperm *LFY* at complementing an *Arabidopsis lfy* mutant [52,56,62]. It is still unknown whether modifications in protein stability, affinity for DNA, sequence-specific recognition or interaction with coregulators account for differences between the two paralogs.

More research is needed to understand the role of *LFY* homologs in gymnosperms, but there is a growing body of evidence suggesting that a minimal network involving LFY and some ABC-type MADS-box proteins was already at work in the reproductive structures of the most recent common ancestor of seed plants. This is of crucial importance because the subsequent rearrangement of such a network after the divergence of gymnosperms is likely to be one of the causative forces of the origin of flowers in angiosperms [63]. Insights into the function of LFY before the evolution of the seed habit can be gained by studying living plants that have retained the free-sporing habit, for example, ferns and mosses.

Review Trends in Plant Science Vol.15 No.6

Back to LEAFY origins: the situation in free-sporing land plants

The *LFY* gene is also present in free-sporing land plant [i.e. lycophytes, ferns and their allies and bryophytes (moss, hornworts and liverworts)] [64,65]. However, the reproductive organs of these plants are so different from flowers that it is difficult to compare the role of LFY in these groups with that in seed plants.

A single functional analysis has been performed in the moss *Physcomitrella patens*. The two close *PpLFY* paralogs are broadly expressed in both the sporophyte (diploid) and the gametophyte (haploid) [65]. When both *PpLFY* genes are mutated, the gametophyte develops normally, whereas the sporophyte is arrested at the first cell division stage right after fertilization. The few sporophytes that do form have general growth defects, suggesting that PpLFY proteins play an important role in the control of cell division during the diploid phase.

Based on the high amino acid conservation of the DNA binding domain, it is reasonable to propose that LFY also functions as a transcription factor in ferns and bryophytes. Indeed, *CrLFY* from the fern *Ceratopteris richardii* can partially complement *Arabidopsis lfy* mutants and the CrLFY protein binds a canonical LFY-binding site CCANT(G/T) *in vitro* [62]. By contrast, PpLFY is inactive in *Arabidopsis* and the protein lacks the capacity of binding this canonical LFY-binding site because an aspartic acid residue (D) replaces a conserved histidine (H) in the DNA binding domain. This feature appears to have been derived in the moss lineage given that the liverwort *Marchantia polymorpha*, which represents a lineage that is sister to all other land plants, resembles vascular plants in possessing the conserved H. It is thus likely that LFY acts as a transcription factor in free-sporing land plants with a slight variation of its mode of action; however, its targets remain unknown. Despite intensive searches, direct orthologs of A, B and C genes have not been identified in bryophytes, ferns and their allies [66–68]. The well-known 'floral' MADS-box genes probably arose later, during the expansion of the MADS family, and are thus specific to seed plants. More studies are needed to identify LFY target genes in early land plants. Such approaches could provide precious information about how LFY fulfilled its role 400 million years ago.

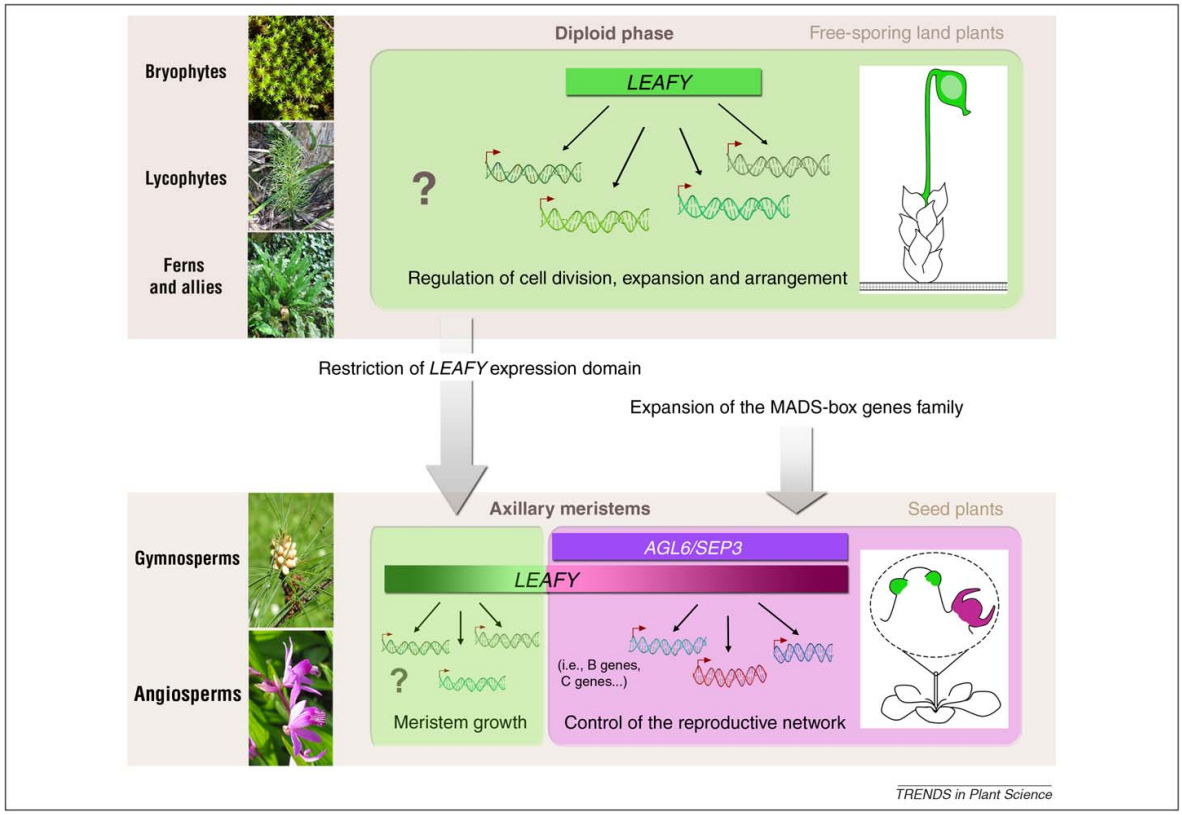


Figure 4. Speculative scenario describing the evolution of LEAFY function. In free-sporing plants such as bryophytes and ferns, *LFY* is broadly expressed throughout the diploid phase (sporophyte, colored in green) where it could act as a regulator of cell divisions, expansion and arrangement (ancestral function) but its targets are still unknown. In seed plants, *LFY* expression is detected at high levels in axillary meristems where it might still exert its control on cell division (young axillary meristems arising from the shoot apex, colored in green). The role of LFY in the inflorescence of grasses or the compound leaves of some legumes indicates that its ancestral function remains active in flowering plants. In the floral meristem of angiosperms (colored in pink), LFY also drives a reproductive network (modern function) that generates the different organs of the flower. This second function involves some 'seed plant specific' MADS box genes both as partners (*SEP3* and/or *AGL6*) or targets (ABC genes) of LFY. Some of the actors of the reproductive network are present in gymnosperms (*LEAFY*, B genes, C genes) and could fulfill a similar function, so a pre-floral version of the reproductive network is likely to exist in this group. However, regulations within this pre-floral network remain to be understood because so far *SEP* genes have only been found in angiosperms, whereas *AGL6*-like genes have also been found in gymnosperms.

The evolution of LEAFY functions: a speculative scenario

In early land plants, *LFY* homologs might have functioned in a minimal network regulating cell division and expansion throughout the sporophyte (Figure 4). *LFY* targets in such plants remain to be identified, but genes involved in cell cycle regulation, growth or differentiation, as well as hormonal signaling, are good candidates.

As vascular plants diversified, *LFY* expression patterns might have undergone a restriction, by changes in *cis*-regulatory elements or in the upstream regulatory network, so that in seed plants *LFY* genes remained inactive in most vegetative tissues and high levels of expression of these genes are first detected in axillary meristems arising on the flank of the shoot apex. Given that meristems are groups of undifferentiated cells that divide intensively, we hypothesize that LFY still exerts its ancestral role on the regulation of cell division in gymnosperms and angiosperms, but in a territory restricted to axillary meristems. This function, which is obvious in the grass inflorescence and more cryptic (but present) in *Arabidopsis* flowers, would have been recruited in some legumes to create compound leaves.

In seed plants, one novel LFY function would be to trigger a reproductive gene network (Figure 4). As shown in model angiosperms, this involves MADS-box transcription factors both as downstream LFY targets (ABC genes) and as LFY partners (AGL6 and/or SEP3). These MADS-box genes are absent from free-sporing plants and probably originated from a diversification of the MADS family in the lineage leading to seed plants. When exactly this modern network appeared remains elusive because its existence in gymnosperms has not been demonstrated. Its emergence probably involved changes in *cis*-elements of recruited targets, to place them under LFY control, as well as the establishment of novel protein–protein interactions. In most living angiosperms, both LFY functions would be present, and evolution has used all the inherent potential of the associated regulatory networks to create the wide variety of inflorescence structures observed in nature.

We are aware that this scenario is speculative, but our aim was to provide a framework that integrated the wealth of recent data. Now that we know more about the phylogeny of extant seed plants, developing new model species in gymnosperms as well as working with several well-placed angiosperm groups (e.g. *Amborella*, Nymphealeas, Piperales, Alismatales, Ranunculids) [69] is key to fully understanding the fascinating history of this peculiar gene and the evolution of the network that regulates floral identity.

Acknowledgements

We thank P. Laufs, M. Frohlich, J. Hofer, M. Blázquez and members of the Koes and Parcy laboratories for discussion, E. Coen and G. Tichtinsky for providing pictures, E. Dorcey for critical reading of the manuscript and the referees for constructive comments. Research in our laboratories is supported by funding from the Centre National de la Recherche Scientifique (CNRS, Action Thématique et Incitative sur Programme, F.P.), the Agence Nationale de la Recherche (ANR, Plant-TFcode, F.P.), the ANR and the Biotechnology and Biological Sciences Research Council (Flower Model, F.P.), and of the Netherlands Organisation for Scientific Research (NWO) to R.K.

References

1 Benlloch, R. *et al.* (2007) Floral initiation and inflorescence architecture: a comparative view. *Ann. Bot. (Lond.)* 100, 659–676
2 Blazquez, M.A. *et al.* (2006) How floral meristems are built. *Plant Mol. Biol.* 60, 855–870
3 Liu, C. *et al.* (2009) Coming into bloom: the specification of floral meristems. *Development* 136, 3379–3391
4 Moyroud, E. *et al.* (2009) The LEAFY floral regulators in Angiosperms: conserved proteins with diverse roles. *J. Plant Biol.* 52, 177–185
5 Wagner, D. (2009) Flower morphogenesis: timing is key. *Dev. Cell* 16, 621–622
6 Irish, V.F. (2010) The flowering of *Arabidopsis* flower development. *Plant J.* 61, 1014–1028
7 Schmid, M. *et al.* (2003) Dissection of floral induction pathways using global expression analysis. *Development* 130, 6001–6012
8 William, D.A. *et al.* (2004) Genomic identification of direct target genes of LEAFY. *Proc. Natl. Acad. Sci. U. S. A.* 101, 1775–1780
9 Hamès, C. *et al.* (2008) Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. *EMBO J.* 27, 2628–2637
10 Liu, C. *et al.* (2009) Regulation of floral patterning by flowering time genes. *Dev. Cell* 16, 711–722
11 Chae, E. *et al.* (2008) An *Arabidopsis* F-box protein acts as a transcriptional co-factor to regulate floral development. *Development* 135, 1235–1245
12 Kodadek, T. *et al.* (2006) Keeping transcriptional activators under control. *Cell* 127, 261–264
13 Simon, R. *et al.* (1994) Fimbriata controls flower development by mediating between meristem and organ identity genes. *Cell* 78, 99–107
14 Lippman, Z.B. *et al.* (2008) The making of a compound inflorescence in tomato and related nightshades. *PLoS Biol.* 6, e288
15 Souer, E. *et al.* (2008) Patterning of inflorescences and flowers by the F-Box protein DOUBLE TOP and the LEAFY homolog ABERRANT LEAF AND FLOWER of Petunia. *Plant Cell* 20, 2033–2048
16 Zhang, S. *et al.* (2003) *Proliferating Floral Organs* (*Pfo*), a *Lotus japonicus* gene required for specifying floral meristem determinacy and organ identity, encodes an F-box protein. *Plant J.* 33, 607–619
17 Bomblies, K. *et al.* (2003) Duplicate *FLORICAULA/LEAFY* homologs *zfl1* and *zfl2* control inflorescence architecture and flower patterning in maize. *Development* 130, 2385–2395
18 Kyozuka, J. *et al.* (1998) Down-regulation of *RFL*, the *FLO/LFY* homolog of rice, accompanied with panicle branch initiation. *Proc. Natl. Acad. Sci. U. S. A.* 95, 1979–1982
19 Rao, N.N. *et al.* (2008) Distinct regulatory role for *RFL*, the rice *LFY* homolog, in determining flowering time and plant architecture. *Proc. Natl. Acad. Sci. U. S. A.* 105, 3646–3651
20 Ikeda, K. *et al.* (2007) Rice *ABERRANT PANICLE ORGANIZATION 1*, encoding an F-box protein, regulates meristem fate. *Plant J.* 51, 1030–1040
21 Ikeda-Kawakatsu, K. *et al.* (2009) Expression level of *ABERRANT PANICLE ORGANIZATION1* determines rice inflorescence form through control of cell proliferation in the meristem. *Plant Physiol.* 150, 736–747
22 Ohmori, S. *et al.* (2009) *MOSAIC FLORAL ORGANS1*, an *AGL6*-like MADS box gene, regulates floral organ identity and meristem fate in rice. *Plant Cell* 21, 3008–3025
23 Becker, A. *et al.* (2005) Floral and vegetative morphogenesis in California poppy (*Eschscholzia californica Cham.*). *Int. J. Plant Sci.* 166, 537–555
24 Prenner, G. *et al.* (2009) The key role of morphology in modelling inflorescence architecture. *Trends Plant Sci.* 14, 302–309
25 Prusinkiewicz, P. *et al.* (2007) Evolution and development of inflorescence architectures. *Science* 316, 1452–1456
26 Parcy, F. (2005) Flowering: a time for integration. *Int. J. Dev. Biol.* 49, 585–593
27 Bradley, D. *et al.* (1997) Inflorescence commitment and architecture in *Arabidopsis*. *Science* 275, 80–83
28 Weigel, D. and Nilsson, O. (1995) A developmental switch sufficient for flower initiation in diverse plants. *Nature* 377, 495–500
29 Kelly, A.J. *et al.* (1995) *NFL*, the tobacco homolog of *FLORICAULA* and *LEAFY*, is transcriptionally expressed in both vegetative and floral meristems. *Plant Cell* 7, 225–234

30 Molinero-Rosales, N. *et al.* (1999) *FALSIFLORA*, the tomato orthologue of *FLORICAULA* and *LEAFY*, controls flowering time and floral meristem identity. *Plant J.* 20, 685–693
31 Souer, E. *et al.* (1998) Genetic control of branching pattern and floral identity during *Petunia* inflorescence development. *Development* 125, 733–742
32 Lee, I. *et al.* (1997) A *LEAFY* co-regulator encoded by *UNUSUAL FLORAL ORGANS*. *Curr. Biol.* 7, 95–104
33 Carroll, S.B. (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134, 25–36
34 Hoekstra, H.E. and Coyne, J.A. (2007) The locus of evolution: evo devo and the genetics of adaptation. *Evolution* 61, 995–1016
35 Pennisi, E. (2008) Evolutionary biology. Deciphering the genetics of evolution. *Science* 321, 760–763
36 Stern, D.L. and Orgogozo, V. (2008) The loci of evolution: how predictable is genetic evolution? *Evolution* 62, 2155–2177
37 Shu, G. *et al.* (2000) LEAFY and the evolution of rosette flowering in violet cress (*Jonopsidium acaule*, *Brassicaceae*). *Am. J. Bot.* 87, 634–641
38 Yoon, H.S. and Baum, D.A. (2004) Transgenic study of parallelism in plant morphological evolution. *Proc. Natl. Acad. Sci. U. S. A.* 101, 6524–6529
39 Sliwinski, M.K. *et al.* (2007) The role of two *LEAFY* paralogs from *Idahoa scapigera* (Brassicaceae) in the evolution of a derived plant architecture. *Plant J.* 51, 211–219
40 Kanrar, S. *et al.* (2008) Regulatory networks that function to specify flower meristems require the function of homeobox genes *PENNYWISE* and *POUND-FOOLISH* in *Arabidopsis*. *Plant J.* 54, 924–937
41 Karim, M.R. *et al.* (2009) A role for *Arabidopsis PUCHI* in floral meristem identity and bract suppression. *Plant Cell* 21, 1360–1372
42 Lee, J. *et al.* (2008) SOC1 translocated to the nucleus by interaction with AGL24 directly regulates *LEAFY*. *Plant J.* 55, 832–843
43 Liu, C. *et al.* (2008) Direct interaction of AGL24 and SOC1 integrates flowering signals in *Arabidopsis*. *Development* 135, 1481–1491
44 Yamaguchi, A. *et al.* (2009) The microRNA-regulated SBP-Box transcription factor SPL3 is a direct upstream activator of *LEAFY*, *FRUITFULL*, and *APETALA1*. *Dev. Cell* 17, 268–278
45 Champagne, C.E. *et al.* (2007) Compound leaf development and evolution in the legumes. *Plant Cell* 19, 3369–3378
46 Hofer, J. *et al.* (1997) *UNIFOLIATA* regulates leaf and flower morphogenesis in pea. *Curr. Biol.* 7, 581–587
47 Taylor, S. *et al.* (2001) *Stamina pistilloida*, the pea ortholog of *Fim* and *UFO*, is required for normal development of flowers, inflorescences, and leaves. *Plant Cell* 13, 31–46
48 Blein, T. *et al.* (2008) A conserved molecular framework for compound leaf development. *Science* 322, 1835–1839
49 Mouradov, A. *et al.* (1998) *NEEDLY*, a *Pinus radiata* ortholog of *FLORICAULA/LEAFY* genes, expressed in both reproductive and vegetative meristems. *Proc. Natl. Acad. Sci. U. S. A.* 95, 6537–6542
50 Frohlich, M.W. and Parker, D.S. (2000) The mostly male theory of flower evolutionary origins: from genes to fossils. *Syst. Bot.* 25, 155–170
51 Carlsbecker, A. *et al.* (2004) The MADS-box gene *DAL1* is a potential mediator of the juvenile-to-adult transition in Norway spruce (*Picea abies*). *Plant J.* 40, 546–557
52 Dornelas, M.C. and Rodriguez, A.P. (2005) A *Floricaula/Leafy* gene homolog is preferentially expressed in developing female cones of the tropical pine *Pinus caribaea* var. *caribaea*. *Genet. Mol. Biol.* 28, 299–307
53 Guo, C.L. *et al.* (2005) Expressions of LEAFY homologous genes in different organs and stages of *Ginkgo biloba*. *Yi Chuan* 27, 241–244
54 Mellerowicz, E.J. *et al.* (1998) *PRFLL* – a *Pinus radiata* homologue of *FLORICAULA* and *LEAFY* is expressed in buds containing vegetative shoot and undifferentiated male cone primordia. *Planta* 206, 619–629
55 Shindo, S. *et al.* (2001) Characterization of a *FLORICAULA/LEAFY* homologue of *Gnetum parvifolium* and its implications for the evolution of reproductive organs in seed plants. *Int. J. Plant Sci.* 162, 1199–1209
56 Shiokawa, T. *et al.* (2008) Isolation and functional analysis of the *CjNdly* gene, a homolog in *Cryptomeria japonica* of *FLORICAULA/LEAFY* genes. *Tree Physiol.* 28, 21–28
57 Vazquez-Lobo, A. *et al.* (2007) Characterization of the expression patterns of *LEAFY/FLORICAULA* and *NEEDLY* orthologs in female and male cones of the conifer genera *Picea*, *Podocarpus*, and *Taxus*: implications for current evo-devo hypotheses for gymnosperms. *Evol. Dev.* 9, 446–459
58 Zahn, L.M. *et al.* (2005) The evolution of the *SEPALLATA* subfamily of MADS-box genes: a preangiosperm origin with multiple duplications throughout angiosperm history. *Genetics* 169, 2209–2223
59 Rijpkema, A.S. *et al.* (2009) The petunia *AGL6* gene has a *SEPALLATA*-like function in floral patterning. *Plant J.* 60, 1–9
60 Theissen, G. and Becker, A. (2004) Gymnosperms orthologs of class B floral homeotic genes and their impact on understanding flower origin. *Crit. Rev. Plant Sci.* 23, 129–148
61 Sundstrom, J. and Engstrom, P. (2002) Conifer reproductive development involves B-type MADS-box genes with distinct and different activities in male organ primordia. *Plant J.* 31, 161–169
62 Maizel, A. *et al.* (2005) The floral regulator LEAFY evolves by substitutions in the DNA binding domain. *Science* 308, 260–263
63 Frohlich, M.W. (2003) An evolutionary scenario for the origin of flowers. *Nat. Rev. Genet.* 4, 559–566
64 Himi, S. *et al.* (2001) Evolution of MADS-box gene induction by *FLO/LFY* genes. *J. Mol. Evol.* 53, 387–393
65 Tanahashi, T. *et al.* (2005) Diversification of gene function: homologs of the floral regulator *FLO/LFY* control the first zygotic cell division in the moss *Physcomitrella patens*. *Development* 132, 1727–1736
66 Münster, T. *et al.* (2002) Evolutionary aspects of MADS-box genes in the eusporangiate fern *Ophioglossum*. *Plant Biol.* 4, 474–483
67 Singer, S.D. *et al.* (2007) Clues about the ancestral roles of plant MADS-box genes from a functional analysis of moss homologues. *Plant Cell Rep.* 26, 1155–1169
68 Tanabe, Y. *et al.* (2003) Characterization of the *Selaginella remotifolia* MADS-box gene. *J. Plant Res.* 116, 71–75
69 Rudall, P.J. *et al.* (2009) Nonflowers near the base of extant angiosperms? Spatiotemporal arrangement of organs in reproductive units of Hydatellaceae and its bearing on the origin of the flower. *Am. J. Bot.* 96, 67–82
70 Smyth, D.R. *et al.* (1990) Early flower development in *Arabidopsis*. *Plant Cell* 2, 755–767
71 Krizek, B.A. and Fletcher, J.C. (2005) Molecular mechanisms of flower development: an armchair guide. *Nat. Rev. Genet.* 6, 688–698
72 Kyozuka, J. *et al.* (2000) Spatially and temporally regulated expression of rice MADS box genes with similarity to *Arabidopsis* class A, B and C genes. *Plant Cell Physiol.* 41, 710–718
73 Yamaguchi, T. and Hirano, H.Y. (2006) Function and diversification of MADS-box genes in rice. *Scientific World J.* 6, 1923–1932
74 Yamaguchi, T. *et al.* (2006) Functional diversification of the two C-class MADS box genes OSMADS3 and OSMADS58 in *Oryza sativa*. *Plant Cell* 18, 15–28

PRESENTATION DES OBJECTIFS

LFY est un facteur de transcription unique chez les plantes terrestres, possédant un domaine de liaison à l'ADN très fortement conservé. De plus, *LFY* s'est très rarement dupliqué, ce qui limite fortement ses possibilités d'évolution. En effet, sa fonction est essentielle chez les angiospermes aussi bien que chez la mousse *Physcomitrella patens,* et donc une modification des propriétes de LFY sans duplication au préalable pourrait avoir des conséquences dramatiques sur le développement de la plante. Pourtant, le rôle de PpLFY1/2 est totalement différent de celui de LFY chez *Arabidopsis thaliana*, et PpLFY1 ne lie pas les mêmes séquences que LFY. Il est possible que PpLFY1 ne soit pas capable de lier l'ADN, mais étant donné le fort degré de conservation du domaine de liaison à l'ADN, il est plus probable qu'il possède une spécificité de liaison divergente. Il semble donc que LFY ait évolué chez *P. patens*, sans pourtant former de famille multigénique.

Les questions auxquelles j'ai tenté de répondre sont les suivantes :

- LFY a-t-il souvent changé de spécificité de liaison à l'ADN lors de l'évolution des plantes terrestres ?
- Comment LFY, facteur de transcription au rôle essentiel, a-t-il pu changer de spécificité pendant l'évolution sans former de famille multigénique ?
- Quelles sont les conséquences d'un tel changement de spécificité sur la régulation des gènes cibles de LFY ?

Au-delà de l'intérêt d'étudier l'évolution de LFY et du réseau floral, cette protéine constitue un modèle de choix pour appréhender l'évolution des facteurs de transcription en général, puisqu'elle est restée unique et facilement identifiable chez l'ensemble des plantes terrestres. La réponse aux questions ci-dessus pourrait donc être utile à l'étude de nombreux facteurs de transcription.

Les objectifs de ma thèse peuvent être divisés en 2 parties :

1) Déterminer l'évolution de la spécificité de liaison de LFY chez les plantes terrestres

Par des approches biochimiques, nous allons tout d'abord déterminer la spécificité de liaison à l'ADN de LFY chez différentes espèces représentatives des grands groupes de plantes terrestres (**chapitre I**). Nous pourrons ainsi préciser la spécificité de liaison de PpLFY1/2, et savoir si LFY a fréquemment changé de spécificité de liaison au cours de l'évolution. Cette spécificité sera ensuite utilisée pour prédire, de manière simple et précise, les sites de liaison à l'ADN de LFY dans un génome, et aboutir ainsi à des conclusions évolutives sur certains gènes régulés par LFY (**chapitre II**). Nous pourrons alors déterminer si les éléments *cis* du réseau ont subi des modifications au cours de l'évolution.

2) Etudier l'évolution du réseau transcriptionnel contrôlé par LFY chez *P. patens*

Une fois précisée la spécificité de liaison de PpLFY1/2 à l'ADN, je chercherai à identifier ses gènes cibles par des approches bioinformatiques et *in planta* (**chapitre III**). Je tenterai également de comprendre comment un tel changement de spécificité a pu avoir lieu au cours de l'évolution, sans que LFY ne forme de famille multigénique. Pour cela, nous allons poser la question de la spécificité ancestrale de LFY. Tout ceci nous permettra d'appréhender les mécanismes contrôlant l'évolution du réseau transcriptionnel orchestré par LFY.

CHAPITRE I :

Evolution de la spécificité de liaison à l'ADN de LFY

chez les plantes terrestres

Dans ce chapitre, j'aborde la question de l'évolution de la spécificité de liaison à l'ADN de LFY. Cette propriété est essentielle chez un facteur de transcription puisqu'elle conditionne sa palette de gènes cibles, et j'ai voulu savoir si elle pouvait varier chez un facteur très conservé, au rôle essentiel et n'ayant pas évolué au sein d'une famille multigénique. D'autres propriétés telles que la multimérisation ou l'interaction avec d'autres partenaires protéiques peuvent également modifier l'action régulatrice d'un facteur de transcription (comme c'est le cas pour les complexes de protéines MADS du modèle ABC) (Immink et al., 2010), mais ces aspects sont moins bien connus pour LFY et donc plus difficiles à étudier, ce qui nous a poussés à nous focaliser sur sa spécificité de reconnaissance de l'ADN.

Pour étudier l'évolution de cette propriété, nous avons choisi une approche biochimique. L'ensemble de ce travail avait été largement initié par Edwige Moyroud, étudiante en thèse. J'ai poursuivi cette approche sur des protéines supplémentaires, et analysé l'ensemble des résultats par la suite.

1) Production de protéines LFY recombinantes

a) Choix des protéines d'intérêt

Chez les angiospermes, nous avons choisi les espèces d'intérêt (**Fig. 7**) en raison de leur position phylogénétique principalement, mais également lorsque l'orthologue de *LFY* semblait montrer un rôle divergent de celui de *LFY*. Les angiospermes peuvent être subdivisées en trois groupes : les monocotylédones, où l'on trouve par exemple les céréales ; les eudicotylédones, qui comprennent plus de 75% des espèces de plantes à fleurs ; et les angiospermes « basales », comprenant entre autres le clade ANA (Amborellacées, Nymphéacées et Austrobaileyales) qui constitue le groupe ayant divergé le plus tôt des autres angiospermes (Moore et al., 2007). Dans le groupe des monocotylédones, nous avons choisi d'étudier *RFL*, l'orthologue de *LFY* chez le riz (*Oryza sativa*). *RFL* est exprimé dans les méristèmes floraux ainsi que dans ceux initiant les branchements de l'inflorescence (Kyozuka et al., 1998), et une lignée *knock-down* par siRNA pour *RFL* montre des défauts de ramification de l'inflorescence et d'émergence des talles (Rao et al., 2008). Ceci suggère que *RFL* n'a pas uniquement un rôle floral mais également architectural, donc potentiellement divergent par rapport à *LFY*. Chez les eudicotylédones et plus particulièrement les Rosidées, nous avons choisi la protéine modèle LFY d'*Arabidopsis thaliana*, et les protéines RoLFY de

la rose (*Rosa chinensis)* et VFL de la vigne (*Vitis vinifera*). *VFL* présente un patron d'expression original puisqu'il est exprimé dans tous les méristèmes, floraux ou végétatifs (Carmona et al., 2002). Enfin, chez les angiospermes basales nous avons choisi la protéine AmboLFY de la plante *Amborella trichopoda*, qui appartient au clade ANA et représente à ce jour l'espèce la plus « basale » des angiospermes (Soltis and Soltis, 2004; Moore et al., 2007).

Nous avons également choisi des protéines LFY et NLY chez les gymnospermes (*Welwitschia mirabilis*, *Picea abies*, *Ginkgo biloba*) et la protéine LFY de la fougère *Ceratopteris richardii*. Enfin, nous nous sommes intéressés à la protéine PpLFY1 de *Physcomitrella patens* puisqu'il a été proposé qu'elle possède une spécificité de liaison à l'ADN différente de celle de LFY (Maizel et al., 2005), ainsi qu'à la protéine MarpoFLO de *Marchantia polymorpha*, pour déterminer l'étendue de ce possible changement de spécificité dans le groupe des mousses.

Espèce	Nom du gène	Numéro du premier acide aminé de la construction	Nom de la construction
Arabidopsis thaliana	*LFY*	40	**LFYΔ ****
		217	**LFY-C ****
Rosa chinensis	*RoLFY*	46	**RoLFYΔ**
		232	**RoLFY-C**
Vitis vinifera	*VFL*	45	**VFLΔ**
Oryza sativa	*RFL*	42	**RFLΔ**
		216	**RFL-C**
Amborella trichopoda	*AmboLFY*	1	**AmboLFY ****
		191	**AmboLFY-C ****
Welwitschia mirabilis	*WelLFY*	59	**WelLFYΔ ****
		247	**WelLFY-C ****
	WelNDLY	57	**WelNDLYΔ ****
		245	**WelNDLY-C ****
Picea abies	*PaNLY*	166	**PaNLY-C**
Ginkgo biloba	*GinLFY*	45	**GinLFYΔ ****
		236	**GinLFY-C ****
Ceratopteris richardii	*CrLFY2*	43	**CrLFY2Δ**
Physcomitrella patens	*PpLFY1*	1	**PpLFY1 ****
Marchantia polymorpha	*MarpoFLO*	inconnu*	**MarpoFLO-C**

Figure 7 : Bilan des différentes constructions protéiques de LFY utilisées dans cette étude. Pour chaque espèce choisie, généralement une protéine entière ou délétée d'environ 40 acides aminés (Δ, en vert), ainsi qu'une protéine comprenant uniquement le domaine C-terminal (-C, en bleu) ont été purifiées. *La séquence entière de *MarpoFLO* n'est pas connue pour l'instant. **Pour ces constructions, les clonages, purifications et SELEX ont été réalisés par d'autres membres de l'équipe (Emmanuel Thévenon, Edwige Moyroud et Renaud Dumas).

b) Purification des protéines

Les protéines LFY ont été surexprimées en système bactérien, puis purifiées par colonne d'affinité de Nickel-Sépharose. En effet, l'ensemble des ADNc ont été clonés dans des vecteurs apportant une étiquette de 6 histidines, en C-terminal ou en N-terminal de la protéine, autorisant ainsi ce type de purification. Cette méthode, rapide et efficace, apporte généralement un degré de pureté suffisant pour les expériences de SELEX que nous effectuons par la suite (**Fig. 8**). Néanmoins, une étape de purification supplémentaire par chromatographie d'exclusion de taille (filtration sur gel) a parfois été effectuée.

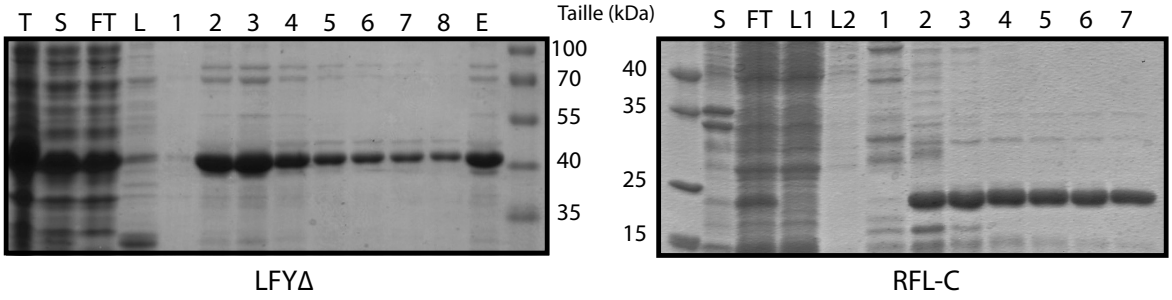


Figure 8 : Exemples de gels d'acrylamide détaillant les fractions récupérées au cours d'une purification sur colonne d'affinité de Nickel-Sépharose. Gel de gauche : purification de LFYΔ (LFY d'*A. thaliana* tronquée de ses 40 premiers acides aminés); gel de droite : purification de RFL-C (orthologue de LFY chez le riz, domaine C-terminal). T : fraction bactérienne totale ; S : fraction soluble après sonication et centrifugation ; FT : flow-through (protéines ayant traversé la colonne sans être retenues) ; L : fraction de lavage (pour RFL-C, deux lavages successifs ont été effectués) ; 1-8 : fractions d'élution de 1,5 mL chacune ; E : fraction protéique finale utilisée pour les expériences. Les tailles d'un marqueur de poids moléculaire sont indiquées en kDa.

2) Expériences de SELEX

Pour déterminer la spécificité de liaison à l'ADN de LFY, nous avons choisi la technique du SELEX (Systematic Evolution of Ligands by EXponential enrichment), qui consiste à sélectionner les séquences reconnues *in vitro* par un facteur de transcription, à partir d'un groupe d'oligonucléotides de séquence aléatoire (Djordjevic, 2007).

La librairie initiale est constituée d'environ 10^{15} à 10^{16} oligonucléotides double brin, composés d'une séquence aléatoire flanquée de bordures constantes de chaque côté. Cette librairie est mise en contact avec la protéine d'intérêt, qui va lier certains oligonucléotides selon sa spécificité de reconnaissance. Par différentes méthodes, on isole les complexes ADN-protéine pour récupérer les oligonucléotides sélectionnés ; ceux-ci sont ensuite amplifiés par PCR, et remis en contact avec la protéine pour renforcer la sélection (**Fig. 9**). La réitération des cycles de sélection permet d'éliminer le bruit de fond de liaison aspécifique.

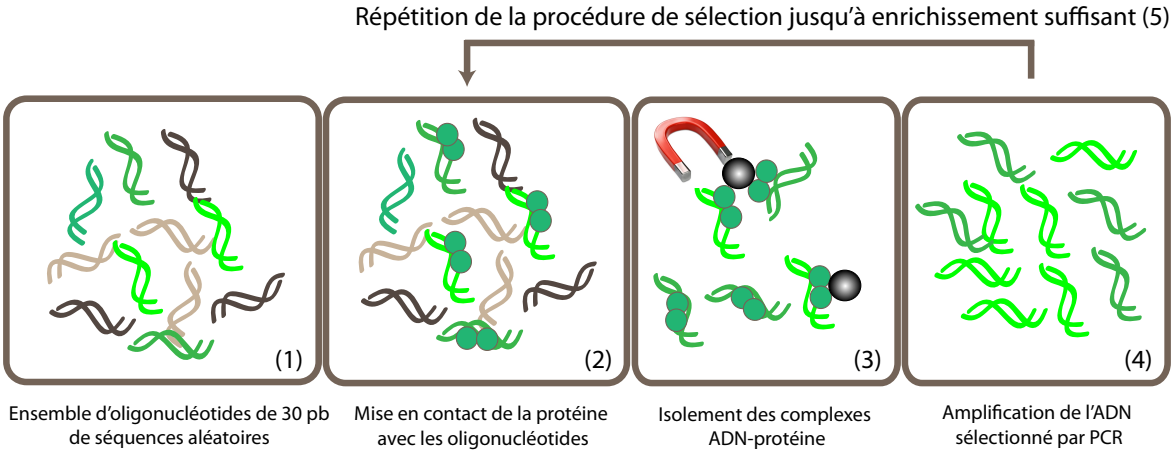


Figure 9 : Principe de la technique de SELEX employée dans cette étude. Une librairie d'oligonucléotides de 73 pb, contenant une séquence aléatoire de 30 pb à leur centre (1), est mise en contact avec la protéine purifiée (en vert, 2). Les complexes ADN-protéine sont isolés grâce à des billes magnétiques recouvertes de nickel (en gris métallique), attirées avec un aimant (3). L'ADN sélectionné est amplifié par PCR (4), puis remis en contact avec la protéine pour effectuer un nouveau cycle de sélection (5).

Nous avons effectué des expériences de SELEX sur l'ensemble des protéines LFY décrites dans la Figure 7. Pour isoler les complexes ADN-protéine, des billes de nickel magnétiques ont été rajoutées à la réaction de liaison, et seront liées par l'étiquette 6-histidines des protéines recombinantes ; l'ensemble billes-protéines-ADN est ensuite isolé grâce à un aimant (**Fig. 9**). Pour vérifier que nous sélectionnons effectivement des oligonucléotides de bonne affinité pour LFY, l'ADN isolé à chaque tour de SELEX est testé en gel retard ; au fur et à mesure des cycles d'enrichissement, on vérifie que la fraction d'ADN complexé à la protéine devient de plus en plus importante (**Fig. 10**).

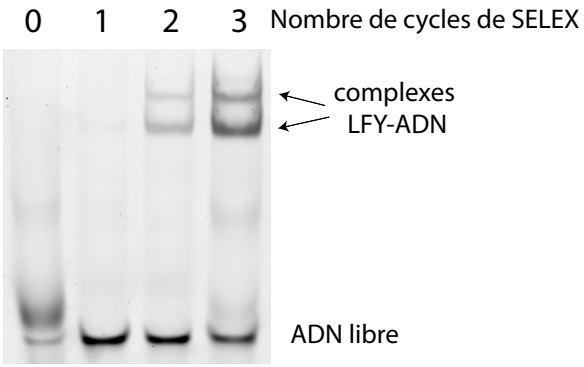


Figure 10 : Exemple d'un gel d'acrylamide reflétant l'enrichissement obtenu au cours d'une expérience de SELEX. La même quantité d'ADN (10 nM) issu de chacun des cycles 0 à 3 du SELEX a été déposée avec 1 µM de protéine. Les retards, représentant les complexes ADN-protéine, s'intensifient au cours de la sélection, alors que la fraction d'ADN libre décroît.

3) Séquençage haut-débit des échantillons de SELEX

a) Des code-barres pour les échantillons

Pour séquencer les oligonucléotides sélectionnés au cours du SELEX, nous avons choisi de recourir au séquençage haut-débit de type Illumina. Cette étape a été réalisée en collaboration avec Norman Warthmann et Markus Schmid (Max Planck Institute for Developmental Biology, Tübingen, Allemagne). Puisque nous disposions de nombreux échantillons différents (un échantillon correspondant à l'ensemble des oligonucléotides sélectionnés à un tour n de SELEX pour une protéine LFY donnée), et que nous n'avions besoin que de quelques milliers de séquences pour chaque échantillon, Edwige Moyroud et moi-même avons développé une stratégie pour pouvoir séquencer tous nos échantillons en une seule fois. Pour cela, à chaque échantillon a été rajouté un « code-barre » (étiquette) de 6 nucléotides, permettant de l'identifier (**Fig. 11**) (Meyer et al., 2007). Tous les échantillons ont ensuite été mélangés pour être séquencés ensemble, en utilisant seulement 10% de la capacité d'une piste de séquençage (le reste étant utilisé pour des échantillons sans rapport avec les nôtres).

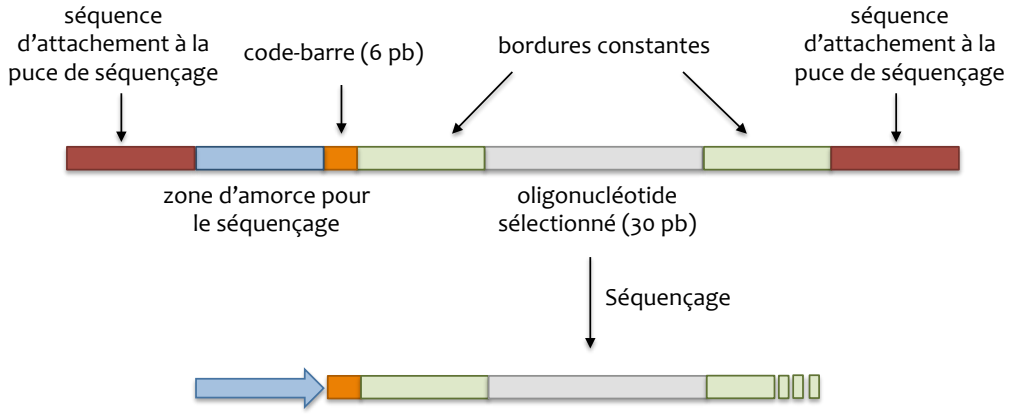


Figure 11 : Organisation des oligonucléotides de SELEX modifiés pour le séquençage Illumina. Les fragments sélectionnés par le SELEX, c'est-à-dire l'oligonucléotide central de 30 pb entouré de ses bordures constantes, sont additionnés d'un « code-barre » (étiquette), d'une zone d'amorce pour le séquençage, et de deux séquences d'attachement à la puce de séquençage de part et d'autre. Le fragment séquencé par Illumina comprendra le code-barre, et l'oligonucléotide (pas nécessairement complet, indiqué par des pointillés) issu du SELEX.

b) Caractéristique des séquences récupérées

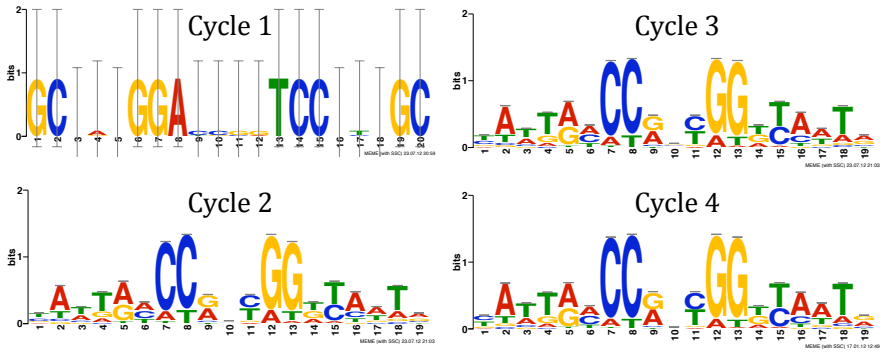
Nous avons obtenu grâce au séquençage environ 4,6 millions de séquences pour l'ensemble des échantillons de SELEX. J'ai créé des programmes en Python pour détecter dans ces séquences l'étiquette de 6 nucléotides et les bordures constantes ; le programme a

ainsi pu récupérer environ 3,7 millions de séquences (de taille strictement supérieure à 10 pb) qui contenaient bien les séquences recherchées. J'ai ensuite regroupé les séquences selon leur étiquette, et j'ai enfin éliminé toutes les séquences répétées, qui peuvent être présentes de très grands nombres de fois. Au total, environ 1,4 millions de séquences uniques ont été isolées, équivalent à en moyenne 46 500 séquences uniques pour chaque échantillon de SELEX. Ce chiffre est néanmoins très variable selon le cycle de SELEX considéré : en effet, le pourcentage de séquences uniques après un cycle de SELEX est généralement autour de 95-99%, alors qu'au quatrième cycle, il peut atteindre 7-8% (**Fig. 12A**). Ceci reflète bien l'enrichissement attendu au cours du processus de sélection, où l'on suppose que les séquences répétées de nombreuses fois après plusieurs cycles de SELEX ont une très bonne affinité pour LFY.

A

Espèce	Construction	Cycle de Selex séquencé	Nombre de séquences totales	Nombre de séquences uniques	Pourcentage de séquences uniques
Arabidopsis thaliana	LFYΔ	1	153491	153280	**99,9**
		2	141680	135295	**95,5**
		3	152259	31803	**20,9**
		4	146783	11058	**7,5**
	LFY-C	1	189261	188940	**99,8**
		2	73201	71448	**97,6**
		3	137800	56324	**40,9**
Physcomitrella patens	PpLFY1	3	138443	131300	**94,8**
		4	98740	51379	**52,0**
		5	103539	8313	**8,0**
Marchantia polymorpha	MarpoFLO-C	1	21789	21358	**98,0**
		2	56297	4537	**8,1**

B



<u>Figure 12 :</u> Analyse des séquences isolées après séquençage de quelques échantillons de SELEX. **A:** Nombre de séquences récupérées après séquençage des échantillons de SELEX de LFYΔ, LFY-C, PpLFY1 et MarpoFLO-C. Pour chaque construction, le nombre de séquences (de taille supérieure à 10 pb) totales et uniques est indiqué pour chaque cycle de SELEX. **B**: Logos obtenus par le programme MEME à partir des 2000 séquences les plus fréquentes des cycles 1, 2, 3 et 4 du SELEX de LFYΔ. La symétrie est imposée au motif, ainsi qu'une taille de 19 pb. Les barres d'erreur sont indiquées sur chaque logo. MEME propose un logo pour les séquences obtenues après le cycle 1, mais celui-ci est issu de l'alignement de 2 séquences uniquement (contre 1644 pour le tour 4), et montre des barres d'erreur très importantes ; un tel motif n'est donc pas significatif.

4) Analyse des résultats de SELEX : spécificité de liaison à l'ADN de LFY chez les plantes terrestres

a) Alignement des séquences issues du SELEX

La détermination de la spécificité de LFY-C a été antérieure aux expériences de SELEX que j'ai menées. Quand je suis arrivée au laboratoire, un SELEX sur LFY-C avait déjà été effectué et séquencé par pyroséquençage 454, et de nombreuses expériences in vitro *avaient permis d'affiner les paramètres d'alignement et de traitement des données pour valider cette spécificité de liaison (voir* ***chapitre II,*** *Article 2*, p. 59*). Ainsi, j'ai pu bénéficier de ces résultats pour améliorer les contraintes d'alignement*.

Pour déterminer la spécificité de liaison de nos protéines à partir des séquences issues du SELEX, il faut utiliser un algorithme qui va rechercher la présence répétée d'un motif dans l'ensemble ou une sous-partie des séquences. L'un des algorithmes le plus fréquemment utilisé est MEME (Multiple EM for Motif Elicitation), qui va chercher à maximiser la vraisemblance d'obtenir un motif donné au fur et à mesure de l'alignement des séquences (Bailey and Elkan, 1994). La spécificité de liaison du facteur de transcription est figurée sous la forme d'un logo, représentant une quantité d'information à chaque position du site de liaison. L'information dérive en quelque sorte de la différence entre la fréquence observée de chaque nucléotide et sa fréquence attendue (25%), avec un maximum d'information à 2 bits (Schneider and Stephens, 1990). Ainsi, une position où chacune des 4 bases sont observées à une fréquence de 25% ne génère aucune information, alors qu'une position où l'une des bases est observée très fréquemment va générer une information proche de 2 bits (positions 7 et 8 du motif de LFYΔ par exemple, **Fig. 12**).

J'ai aligné des séquences non-répétées pour constituer les motifs, car pour LFY-C, l'alignement des 2500 séquences (répétées ou non) issues du SELEX générait un motif qui décrivait mal la liaison *in vitro* de la protéine, alors qu'en utilisant les 494 séquences uniques le motif était de bien meilleure qualité. Les séquences de très bonne affinité pour LFY peuvent en effet apparaître un grand nombre de fois pendant le séquençage, notamment après plusieurs cycles de SELEX (la même séquence peut être retrouvée plus de 35 000 fois dans les échantillons que nous avons séquencés), ce qui va orienter le motif de liaison vers une séquence « cul-de-sac ». Pour tenir compte néanmoins de cette répétition, nous avons choisi de trier les séquences d'après leur occurrence et de n'aligner, pour chaque échantillon, que les

premières séquences uniques ; ainsi les séquences répétées gardent un poids important dans la constitution du motif. Les algorithmes tels que MEME demandent énormément de temps pour aligner de grands nombres de séquences (aligner 10 000 séquences prendrait approximativement une semaine). Pour cela, nous avons décidé d'utiliser uniquement une sous-partie des séquences récupérées lors du séquençage et d'en soumettre 2000 à MEME ; le programme aligne ensuite une sous-partie de ces séquences pour identifier un motif conservé.

L'alignement a été réalisé avec les paramètres par défaut de MEME, sauf lorsque la spécificité de la protéine semblait similaire à celle de LFY-C. Dans ce cas, par analogie avec LFY-C, certaines contraintes ont été appliquées à l'alignement : la symétrie du motif, puisque LFY-C lie l'ADN sous forme de dimère, et une taille de 19 pb, ce qui correspond à la longueur des contacts entre LFY-C et l'ADN (Hamès et al., 2008). Ainsi, chaque moitié du motif de LFY-C correspond à la liaison d'un monomère sur l'ADN.

b) Evolution de la spécificité de liaison de LFY à l'ADN chez les plantes terrestres

Les logos fournis par MEME à partir de l'alignement des 2000 premières séquences de chaque SELEX sont présentés sur la figure 13. LFY-C montre un motif avec une information forte sur les positions 7/8 et 12/13, et quasiment nulle au centre du motif. Cette spécificité de liaison est semblable à celle déterminée précédemment grâce à l'alignement des 494 séquences uniques de SELEX (**Article 2**, p. 59). Les motifs obtenus pour les autres protéines LFY nous ont permis d'aboutir à de nombreuses conclusions :

- **LFYΔ présente une spécificité de liaison similaire à LFY-C, c'est donc majoritairement un seul dimère de la partie C-terminale de LFY qui contrôle la reconnaissance spécifique des bases de l'ADN.** Pourtant, la protéine LFY entière est capable de former des complexes d'ordre supérieur à un dimère (visibles en gel retard), et nous nous demandions si ces complexes pouvaient contacter l'ADN ou imposer des contraintes structurales sur une plus grande distance que 19 pb ; le résultat obtenu nous montre que ce n'est pas le cas et qu'il est valide d'étudier la spécificité de LFY en se basant sur le domaine C-terminal. De manière générale, les domaines de liaison à l'ADN sont souvent utilisés pour justifier d'une reconnaissance spécifique de l'ADN, mais la participation des autres domaines protéiques à cette spécificité est très rarement évaluée.

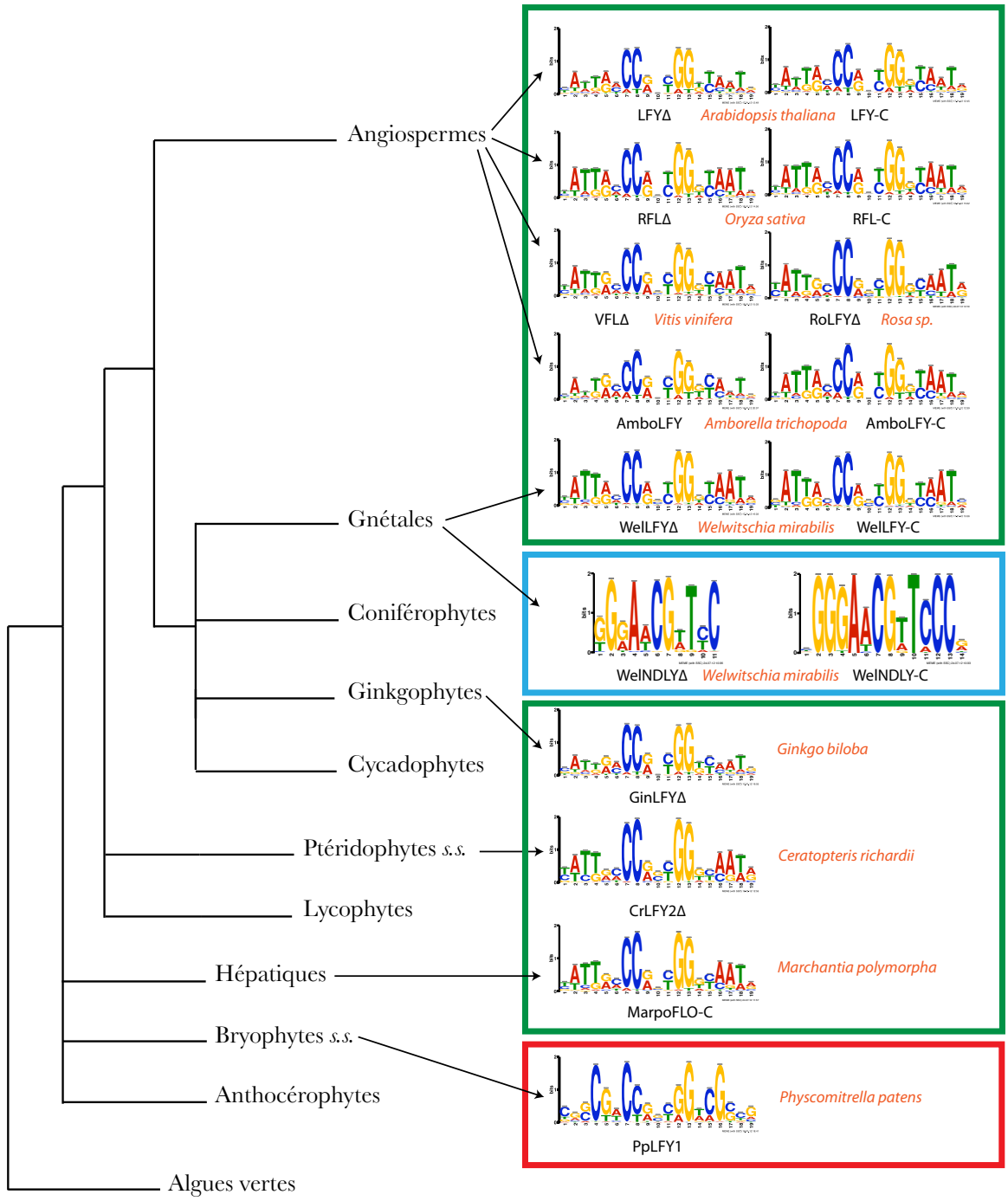


Figure 13: Logos issus de l'alignement des 2000 séquences uniques les plus abondantes de chacun des SELEX réalisés, replacés sur un arbre phylogénétique simplifié des plantes terrestres. Les logos sont obtenus grâce au logiciel MEME, en imposant un motif symétrique de 19 pb, sauf pour les SELEX effectués sur les protéines NLY où aucune contrainte n'est imposée. Les logos obtenus pour LFY chez l'ensemble des angiospermes, gymnospermes, fougères, et hépatiques (encadrés en vert) sont très similaires. Le logo de PpLFY1 (encadré en rouge) a divergé. Les logos obtenus pour les protéines NLY de *Welwitschia mirabilis* (encadrés en bleu) sont représentés sur cette figure, bien que leur spécificité de liaison n'ait pas pu être confirmée *in vitro*. Pour PaNLY-C et GinLFY-C, l'alignement engendre un motif très long qui ne correspond pas à la spécificité de liaison des protéines, et n'a donc pas été représenté sur cette figure.

- **Des hépatiques (*Marchantia polymorpha*) aux angiospermes, LFY présente une spécificité de liaison quasiment inchangée.** Il doit donc exister une forte pression de sélection pour que cette spécificité ait été conservée pendant plus de 400 millions d'années d'évolution (Clarke et al., 2011), soulignant le rôle essentiel de LFY et de son réseau de régulation chez les plantes terrestres.

- **Malgré cette forte conservation de la spécificité de liaison de LFY, il existe deux cas où elle a divergé.** Les protéines WelNDLY-C et WelNDLYΔ de *Welwitschia mirabilis* montrent un motif de 11 à 14 pb, d'allure symétrique, avec un fort contenu en information sur la majorité des positions. Pourtant, ni WelNDLY-C ni WelNDLYΔ ne lient correctement les séquences de meilleure affinité prédite par ce motif en gel retard. WelNDLYΔ et WelLFYΔ montrent pourtant bien des différences de spécificité de liaison *in vitro* (voir **Article 4**, p. 88) : ces deux protéines ont bien des comportements distincts. Nous ne comprenons pas pour l'instant le motif obtenu par SELEX, mais on peut imaginer que les protéines NLY ont une spécificité de liaison complexe qui dépend par exemple des conditions expérimentales, et qui serait alors différente en SELEX et en gel retard.

- **Le second cas de divergence de spécificité observé est celui de PpLFY1**, qui présente un motif d'allure symétrique (même sans contrainte de symétrie) et ressemblant dans son organisation à celui de LFYΔ. En effet, le cœur du motif (positions 7 à 13) est similaire à celui de LFYΔ, mais les bordures apparaissent différentes, notamment au niveau des positions 4 et 16. **Cette différence de spécificité explique très certainement la différence de comportement observée entre PpLFY1 et le reste des protéines LFY** (Maizel et al., 2005). Les raisons moléculaires de ce changement de spécificité et son impact sur la régulation des gènes cibles du réseau seront évalués dans le chapitre III des résultats.

L'évolution du facteur de transcription LFY est donc complexe : il existe une très forte conservation de sa spécificité de liaison chez la quasi-totalité des plantes terrestres, mais elle a pourtant divergé dans certains cas. Le réseau transcriptionnel contrôlé par LFY a donc parfois varié en trans*, ce qui n'exclut évidemment pas qu'il puisse varier en* cis*. Peut-on détecter la présence de ces changements* cis *chez des gènes régulés par LFY?*

CHAPITRE II :

Prédiction de la liaison de LFY à ses gènes cibles

Nous avons vu dans le chapitre précédent qu'à l'exception de deux cas, LFY avait gardé une spécificité de liaison à l'ADN très semblable chez l'ensemble des plantes terrestres. Peut-on utiliser cette information pour prédire la position des sites de liaison de LFY dans le génome de plantes très diverses? Pouvons-nous ainsi étudier l'évolution des gènes régulés par LFY ? Nous allons voir qu'en utilisant les motifs de liaison de LFY obtenus grâce au SELEX, nous pouvons construire un modèle prédisant de manière sensible et spécifique le positionnement de ses sites de liaison, ce qui nous permet de commencer à appréhender l'évolution des éléments *cis* du réseau.

1) Prédire la liaison de LFY à l'ADN

a) Matrices de fréquence, matrices de poids

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	0,07	0,73	0,24	0,04	0,52	0,37	0,12	0,00	0,48	0,18	0,04	0,10	0,02	0,19	0,03	0,66	0,59	0,16	0,42
C	0,36	0,07	0,12	0,03	0,01	0,42	0,86	0,90	0,02	0,32	0,47	0,00	0,01	0,02	0,44	0,27	0,06	0,05	0,16
G	0,16	0,05	0,06	0,27	0,44	0,02	0,01	0,00	0,47	0,32	0,02	0,90	0,86	0,42	0,01	0,03	0,12	0,07	0,36
T	0,42	0,16	0,59	0,66	0,03	0,19	0,02	0,10	0,04	0,18	0,48	0,00	0,12	0,37	0,52	0,04	0,24	0,73	0,07

Matrice de fréquences (f)

$$(1)\quad p_{(i,j)} = \ln\left(\frac{f_{(i,j)}}{f_{(i,\max(A,C,G,T))}}\right)$$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	-1,84	0,00	-0,89	-2,73	0,00	-0,13	-1,97	-5,63	0,00	-0,57	-2,60	-2,22	-3,81	-0,77	-2,99	0,00	0,00	-1,52	0,00
C	-0,16	-2,41	-1,60	-3,26	-4,40	0,00	0,00	0,00	-3,24	0,00	-0,03	-5,89	-5,01	-3,31	-0,17	-0,88	-2,35	-2,76	-0,97
G	-0,97	-2,76	-2,35	-0,88	-0,17	-3,31	-5,01	-5,88	-0,03	0,00	-3,24	0,00	0,00	0,00	-4,40	-3,26	-1,60	-2,41	-0,16
T	0,00	-1,52	0,00	0,00	-2,99	-0,77	-3,81	-2,22	-2,60	-0,57	0,00	-5,63	-1,97	-0,13	0,00	-2,73	-0,89	0,00	-1,84

Matrice de poids (p)

$$(2)\quad S = \sum_{i=1}^{19} p_{(i,n_i)}$$

Calcul du score (S) de séquences individuelles

TATTACCCACTGGGTAATA — S = 0 (score maximum)

AGGCCGGGCTGCCCGGCCT — S = -64,7 (score minimum)

(scores intermédiaires)

Figure 14 : Conversion d'une matrice de fréquences en matrice de poids, et calcul du score d'une séquence donnée. La matrice de fréquence contient les fréquences observées de chaque nucléotide j en position i lors de l'alignement effectué par MEME. Chaque fréquence est convertie en poids grâce à l'équation (1), où $f_{(i,\max(A,C,G,T))}$ est la fréquence la plus forte à la position i. Chaque séquence se voit attribuer un score correspondant à la somme des poids des nucléotides qui la constituent, selon l'équation (2), où n_i représente le nucléotide de la séquence à la position i. Toutes les séquences de 19 pb possibles possèdent donc un score, intermédiaire entre le score maximum 0 et le score minimum, ici de -64,7 (les valeurs de poids correspondantes à cette séquence sont surlignées en couleur sur la matrice de poids).

A partir des alignements que nous avons obtenus par le SELEX, nous pouvons construire un outil prédisant le positionnement des sites de liaison de LFY sur l'ADN. En effet, l'alignement nous fournit une matrice de fréquences, dont les valeurs sont les fréquences observées de chaque nucléotide à chaque position du site de liaison. Cette matrice peut être convertie en une matrice poids/position (fréquemment appelée Position Weight Matrix ou PWM), par une transformation de type logarithmique qui va convertir les fréquences en pénalités, d'autant plus fortes que la fréquence est faible (**Fig. 14)** (Wasserman, 2004; Stormo, 2000). Ainsi, la base la plus fréquente à une position donnée aura un poids de 0, alors qu'une base peu fréquente aura un poids très négatif. L'intérêt d'une telle conversion est que les poids individuels de chaque nucléotide peuvent être additionnés pour calculer le score de n'importe quelle séquence (de la longueur du motif de liaison) (**Fig. 14**). Or il a été montré que le score d'une séquence est proportionnel au logarithme de son affinité (K_D) pour le facteur de transcription (Stormo, 2000; Liu and Clarke, 2002). Ainsi, grâce à la description précise de la spécificité de liaison de LFY, on peut théoriquement prédire de façon très simple son affinité pour n'importe quelle séquence d'ADN de 19 pb.

b) Force prédictive du modèle chez *Arabidopsis thaliana*

Nous avons utilisé le modèle de calcul de scores pour LFY-C afin de prédire ses sites de liaison dans le génome d'*A. thaliana*. Ceci a abouti à la publication d'un article (**Article 2**, p. 59) auquel j'ai participé. Je vais résumer ici les conclusions principales de cet article concernant le pouvoir prédictif de ce modèle de liaison.

Pour évaluer dans quelle mesure le calcul de scores prédisait correctement les sites de liaison de LFY-C *in vitro*, nous avons comparé, pour un ensemble d'oligonucléotides, leurs scores prédits par la matrice de poids avec leurs scores expérimentaux, mesurés par gel retard (Man and Stormo, 2001) (voir **chapitre III** pour des détails sur la technique de mesure par gel retard). Nous avons obtenu une très bonne corrélation entre les données prédites et expérimentales (r^2 =0,81), en considérant notamment que les positions du site de liaison de LFY-C ne contribuaient pas de façon indépendante au calcul de score (**Figure 1**). Pour savoir si ce modèle était également prédictif *in vivo*, nous avons effectué une expérience de ChIP-Seq pour LFY et nous avons comparé les positions des sites de liaison de LFY prédits à celles obtenues par le ChIP-Seq. Au niveau de nombreuses régions de liaison de LFY (situées à proximité des gènes *AP1* ou *TFL1* par exemple), l'adéquation entre le profil de scores (valeurs de scores sur toutes les séquences de 19 pb glissantes d'un grand fragment) et le

profil du pic de ChIP-Seq était très bonne (**Figure 4**). A l'échelle génomique, pour un seuil de score donné, jusqu'à 25% des sites prédits par notre modèle correspondaient effectivement à des sites liés par LFY *in vivo* (**Supplemental Table 3**). Bien que le nombre de faux positifs reste important, un tel pouvoir prédictif est néanmoins très élevé puisque le modèle utilise uniquement l'information d'une séquence génomique pour fournir ces prédictions : de manière générale, ce modèle s'avère hautement prédictif pour détecter les sites de liaison de LFY-C *in vitro* et *in vivo*. Nous disposons ainsi d'un outil rapide et simple d'utilisation pour rechercher les sites de liaison de LFY dans une séquence génomique.

L'expérience de ChIP-Seq nous a également permis d'analyser la liste des gènes liés par LFY *in planta*. Parmi ceux-ci, se retrouvent des gènes cibles de LFY comme *AP1*, *TFL1*, *AG* ou encore *SEP3* (**Tableau 1**), mais surtout de nombreux gènes pour lesquels aucun indice d'une possible régulation par LFY n'existe. Ceci ouvre un important champ d'étude pour déterminer si ces gènes sont effectivement régulés par LFY, et comprendre le rôle biologique de ces régulations qui pourraient révéler de nouvelles fonctions pour LFY.

Notre expertise technique sur la biochimie de LFY m'a permis de participer à une autre étude concernant les gènes cibles de LFY chez *A. thaliana* (**Article 3**, p. 74). Ma contribution à cet article est exclusivement expérimentale, et comprend la réalisation de gels retards (**Figure 4F**). Néanmoins, il est intéressant de mettre en relation cet article, où un ChIP-chip (ChIP suivi d'une hybridation sur micro-array) sur LFY a été effectué, avec la publication de notre équipe. Le motif obtenu par le ChIP-chip de LFY est très similaire à celui que nous avons obtenu en SELEX et en ChIP-Seq, ce qui confirme la cohérence de chacune de ces techniques expérimentales (**Figure 4A**). Le ChIP-chip a permis d'identifier un sous-motif de liaison de LFY (appelé motif secondaire dans la publication, mais en réalité compris dans le modèle biophysique que nous avons construit dans l'Article 2), particulièrement présent dans les gènes liés par LFY participant aux réponses aux stress biotiques (**Figure 4B**). J'ai pu montrer que LFY liait effectivement des oligonucléotides contenant ce motif *in vitro* (**Figure 4F**). Des études *in planta* associées (**Figure 3**) suggèrent que LFY inhibe les mécanismes de réponse aux stress biotiques lors de la floraison, délocalisant ainsi les ressources disponibles vers la reproduction, aux dépens de la défense contre les pathogènes. Les gènes cibles identifiés par le ChIP-chip ont ici permis d'appréhender une nouvelle fonction pour LFY, totalement inattendue.

2) Prédiction de la régulation des gènes floraux par LFY chez les angiospermes et les gymnospermes

Nous avons appliqué le modèle prédictif établi pour LFY-C à l'étude de l'évolution de la régulation de quelques gènes par LFY. En effet, le SELEX nous a fourni la spécificité de liaison de nombreux orthologues de LFY, pour lesquels nous pouvons donc construire un modèle similaire. Dans les autres cas, nous pouvons utiliser le modèle construit pour LFY-C puisque la spécificité de LFY est similaire chez l'ensemble des angiospermes et des gymnospermes. Ceci peut nous permettre d'inférer l'existence de régulations par LFY chez des plantes non modèles, en utilisant uniquement une information génomique.

a) Evolution de la régulation des gènes de la famille d'*AGAMOUS* par LFY chez les angiospermes

Nous avons utilisé le modèle prédictif développé pour LFY-C pour étudier les sites de liaison de LFY dans le plus grand intron des gènes de la famille d'*AG* chez différentes espèces d'angiospermes (**Article 2**, p. 59). Pour cela, nous avons calculé une valeur appelée l'Occupation Prédite (POcc pour Predicted Occupancy) pour ces introns, qui permet d'intégrer l'information de tous les scores successifs de 19 pb sur une longueur donnée ; la POcc représentera alors le nombre de molécules LFY liées à la séquence entière (Liu and Clarke, 2002; Granek and Clarke, 2005; Roider et al., 2007). Nous avons constaté que les valeurs de POcc étaient toujours plus fortes lorsque le gène analysé était régulé par LFY ou lorsque les données expérimentales le suggéraient. Ainsi chez *A. thaliana*, la POcc d'*AG* est plus élevée que celle de *SHATTERPROOF 1* et *2* (*SHP*) et *SEEDSTICK* (*STK*), gènes impliqués dans des aspects tardifs du développement du carpelle, du fruit et de la graine, et non régulés par LFY (**Figure 5C,** voir aussi **Fig. 15**). Puisqu'une forte valeur de POcc est également observée pour les orthologues d'*AG* chez les céréales (**Figure 5B**), nous avons pu proposer que la liaison d'*AG* ou *SHP* par LFY existait déjà avant la divergence entre eudicotylédones et monocotylédones (**Figure 5A**). Ceci est fortement soutenu par de nombreux indices expérimentaux, mais nous avons pu le mettre en évidence en utilisant uniquement des séquences génomiques.

Pourtant, pour les introns des orthologues d'*AG* ou *SHP* présentant une forte valeur de POcc, le profil de scores était très variable, et ni la séquence ni le score des meilleurs sites de liaison de LFY n'apparaissaient conservés (**Figure 6**). Ces variations ne signifient pas

forcément que la liaison de l'élément *cis* sera abolie, puisque le profil de scores (traduit par la valeur de POcc) est toujours propice à une liaison par LFY. Il existe donc une grande fluidité au niveau des sites de liaison, mais qui n'implique peut-être pas forcément une modification de la liaison par LFY. De façon intéressante, la mesure de la valeur de POcc permet de dépasser cette fluidité pour intégrer l'information du paysage de scores d'un fragment donné.

De manière plus générale, cette étude a révélé que, même si la spécificité de liaison de LFY à l'ADN était conservée chez les angiospermes, les éléments *cis* au niveau d'un gène donné ne l'étaient pas forcément. Nous disposons donc d'un premier exemple de modifications du réseau en *cis* sans modification en *trans*, même si l'impact de ces modifications en *cis* sur la régulation des gènes concernés reste à déterminer.

b) Liaison de LFY aux gènes *AGAMOUS* et *SHATTERPROOF* chez les Rosacées

En collaboration avec l'équipe de Mohammed Bendahmane (Laboratoire de Reproduction et Développement des Plantes, ENS de Lyon), nous nous sommes à nouveau penchés sur la question de la régulation d'*AG* et *SHP* par LFY, mais plus particulièrement chez les Rosacées. Chez *A. thaliana*, LFY régule *AG* pour l'accomplissement de la fonction C, mais pas *SHP* qui est impliqué plus tardivement dans le développement du gynécée, du fruit et de l'ovule (Parcy et al., 1998; Pinyopich et al., 2003). Chez *Antirrhinum majus*, l'orthologue d'*AG* (*FARINELLI*) et celui de *SHP* (*PLENA*) se partagent la fonction C, *PLENA* contrôlant l'identité des carpelles et des étamines alors que *FARINELLI* a un rôle restreint à l'identité des étamines (Bradley et al., 1993; Davies et al., 1999). *PLENA*, qui est pourtant l'orthologue de *SHP*, est régulé par LFY, alors que *FARINELLI* ne semble pas l'être (Causier et al., 2005; Causier et al., 2009). La régulation d'*AG/SHP* par LFY et le contrôle de la fonction C ont donc varié chez les angiospermes, et nous nous sommes intéressés à cette situation chez les Rosacées (**Fig. 15**).

La surexpression des orthologues d'*AG* ou *SHP* de la rose (*RoAG/MASAKO C1* ou *RoSHP/MASAKO D1*) chez *A. thaliana* conduisent à un phénotype caractéristique de la surexpression d'un gène C : conversion des sépales en carpelles et des pétales en étamines (Kitahara et al., 2004), suggérant que *RoAG* et *RoSHP* sont tous les deux capables de remplir la fonction C *in planta*. Il est donc possible que, chez les Rosacées, la fonction C soit contrôlée par *RoAG* et *RoSHP* de façon redondante (**Fig. 15**). Nous avons cherché à

comprendre le rôle de LFY dans la régulation de *RoAG* et *RoSHP* chez les Rosacées, et le lien de ces gènes avec l'accomplissement de la fonction C.

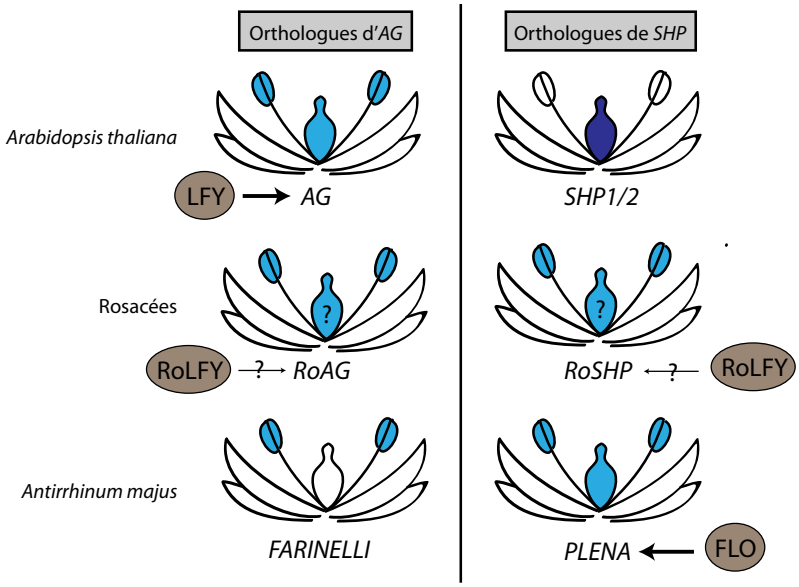
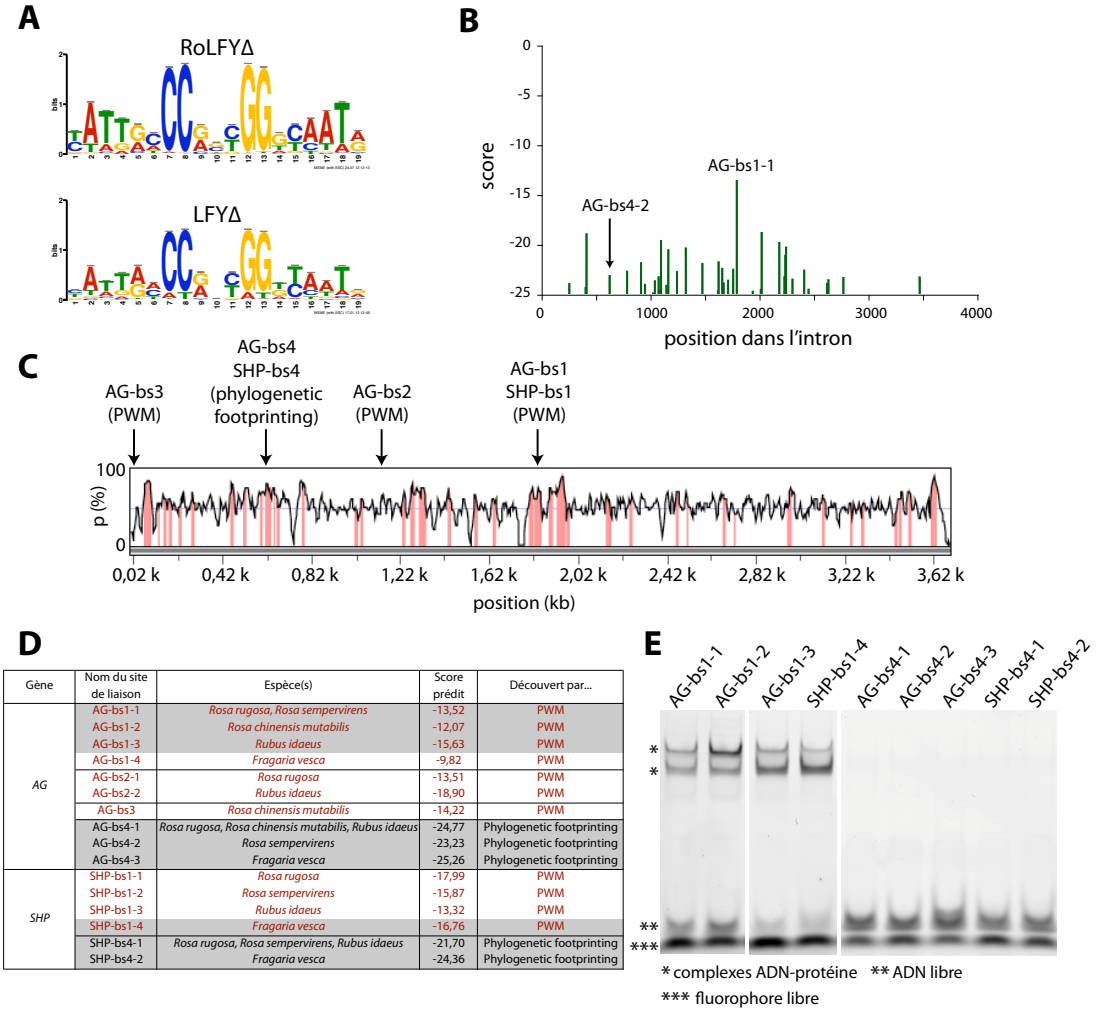


Figure 15 : Evolution de la fonction d'*AG* et *SHP* et de leurs orthologues chez *Arabidopsis thaliana* et *Antirrhinum majus*, et hypothèses sur ces gènes chez les Rosacées. Dans les fleurs schématiques possédant sépales, pétales, étamines et pistil, un organe colorié en bleu signifie que le gène associé a un rôle dans l'identité de cet organe, alors qu'un organe colorié en bleu foncé signifie que le gène a un rôle plus tardif dans le développement de cet organe. La régulation d'un gène par LFY ou son orthologue est indiquée par une flèche lorsqu'elle a été démontrée. Chez les Rosacées, la situation reste pour le moment hypothétique. Inspiré de (Causier et al., 2005).

Pour cela, un SELEX sur RoLFYΔ (orthologue de LFY chez *Rosa chinensis*, la protéine est tronquée de ses 46 premiers acides aminés) a été effectué ; la spécificité de liaison obtenue est très proche de celle de LFY d'*A. thaliana*, comme pour toutes les protéines LFY d'angiospermes (**Fig. 16A**). Le plus grand intron des orthologues d'*AG* et de *SHP* a été isolé et séquencé chez diverses espèces de Rosacées (*Rosa rugosa*, *Rosa sempervirens*, *Rosa chinensis mutabilis*, *Rubus idaeus* et *Fragaria vesca*). En utilisant la matrice poids/position issue du SELEX de RoLFYΔ, nous avons pu identifier un ou deux sites de liaison de RoLFYΔ dans chacun de ces introns (**Fig. 16B, C et D**), dont certains ont été testés et validés en gel retard (**Fig. 16E**). Le site retrouvé dans chacun des introns (AG-bs1 et SHP-bs1) se situe dans des zones de forte conservation locale de séquence (**Fig. 16 C**), mais la séquence et le score de l'élément de liaison varient pourtant légèrement d'une espèce à l'autre. Il existe donc une fluidité, à nouveau, des sites de liaison de LFY, même chez des espèces très proches appartenant à la même famille.

Un site potentiel de liaison de LFY avait été précédemment identifié par phylogenetic footprinting dans le second intron d'*AG* chez les Brassicacées, en utilisant le consensus de liaison précédemment établi pour LFY (CCANTGT/G) (Hong et al., 2003). Ce même site

peut être retrouvé dans les introns des orthologues d'*AG* et *SHP* des Rosacées (AG-bs4 et SHP-bs4) mais les séquences situées à cette position possèdent en réalité un mauvais score de liaison pour RoLFYΔ et ne sont pas reconnus par la protéine *in vitro* (**Fig. 16E**). Nous proposons donc qu'une information de conservation de séquence n'est pas toujours suffisante pour détecter les sites de liaison d'un facteur de transcription, et que l'utilisation d'un modèle performant de prédiction des sites de liaison peut se révéler beaucoup plus efficace.



Gène	Nom du site de liaison	Espèce(s)	Score prédit	Découvert par...
AG	AG-bs1-1	*Rosa rugosa, Rosa sempervirens*	-13,52	PWM
	AG-bs1-2	*Rosa chinensis mutabilis*	-12,07	PWM
	AG-bs1-3	*Rubus idaeus*	-15,63	PWM
	AG-bs1-4	*Fragaria vesca*	-9,82	PWM
	AG-bs2-1	*Rosa rugosa*	-13,51	PWM
	AG-bs2-2	*Rubus idaeus*	-18,90	PWM
	AG-bs3	*Rosa chinensis mutabilis*	-14,22	PWM
	AG-bs4-1	*Rosa rugosa, Rosa chinensis mutabilis, Rubus idaeus*	-24,77	Phylogenetic footprinting
	AG-bs4-2	*Rosa sempervirens*	-23,23	Phylogenetic footprinting
	AG-bs4-3	*Fragaria vesca*	-25,26	Phylogenetic footprinting
SHP	SHP-bs1-1	*Rosa rugosa*	-17,99	PWM
	SHP-bs1-2	*Rosa sempervirens*	-15,87	PWM
	SHP-bs1-3	*Rubus idaeus*	-13,32	PWM
	SHP-bs1-4	*Fragaria vesca*	-16,76	PWM
	SHP-bs4-1	*Rosa rugosa, Rosa sempervirens, Rubus idaeus*	-21,70	Phylogenetic footprinting
	SHP-bs4-2	*Fragaria vesca*	-24,36	Phylogenetic footprinting

Figure 16 : Prédiction et validation de sites de liaison de RoLFYΔ dans les introns d'orthologues d'*AG* ou *SHP* de différentes espèces de Rosacées. A : Comparaison des logos obtenus par le SELEX de RoLFYΔ et de LFYΔ. Les 2000 séquences uniques les plus fréquentes du SELEX ont été alignées grâce au logiciel MEME. **B :** Exemple du profil de score de l'orthologue d'*AG* de *Rosa sempervirens*, et position du meilleur site de liaison prédit par la matrice de RoLFYΔ (AG-bs1-1) ou du site de liaison prédit par phylogenetic footprinting (AG-bs4-2). **C :** Pourcentage de conservation de séquence (p) des introns de l'ensemble des orthologues d'*AG* et de *SHP* isolés (calculé grâce au logiciel mVista). La fenêtre glissante de calcul de conservation de séquence est de 20 pb, et les zones coloriées en rose sont conservées à plus de 70% chez l'ensemble des introns. La position des différents sites de liaison obtenus grâce à la matrice (PWM) ou bien grâce au phylogenetic footprinting est indiquée. **D :** Bilan des sites de liaison identifiés dans les introns des orthologues d'*AG* et *SHP*. PWM : Position Weight Matrix, matrice poids/position de RoLFYΔ. Les sites de liaison qui ont été testés en gel retard sont surlignés en gris. **E :** Validation par gel retard de certains sites de liaison de RoLFY, chaque piste contient 1 µM de protéine recombinante RoLFYΔ.

D'autres expériences sont en cours dans le laboratoire de M. Bendahmane pour valider la régulation d'*AG/SHP* par RoLFY : des hybridations *in situ* chez *Rosa gallica* montrent que *RoLFY* est exprimé précocement dans le méristème floral, puis dans chacun des organes floraux émergents, de façon similaire à l'expression de *LFY* chez *A. thaliana*. L'expression de *RoLFY* précède donc celle de *RoAG* et *RoSHP*, qui sont tous les deux exprimés de manière précoce dans les carpelles et les étamines. En comparaison avec *SHP* chez *A. thaliana*, *RoSHP* est donc exprimé beaucoup plus tôt lors du développement du carpelle, ce qui suggère qu'il ait un rôle dans son identité ou sa mise en place et pas uniquement dans son développement tardif. D'autre part, la complémentation d'un mutant *lfy-12* (mutant *lfy* nul chez *A. thaliana*) par *pLFY::RoLFY* (ADNc de *RoLFY* sous le contrôle du promoteur de *LFY* d'*A. thaliana*) rétablit l'identité des étamines et des carpelles (même s'il existe des défauts d'indétermination de la fleur et de formation des pétales). La protéine RoLFY, exprimée dans un territoire adéquat, est donc capable d'activer, entre autres, le gène de fonction C. Pour savoir si cette activation est également possible chez la Rose, des expériences d'agro-infiltration d'une construction *35S::RoLFY-VP16* dans des pétales de rose sont en cours. En effet, LFY seul ne peut pas activer la transcription génique chez la levure, mais ceci est rétabli s'il est fusionné au domaine activateur de la transcription VP16 (Parcy et al., 1998).

La combinaison de toutes ces expériences pourra déterminer si, *in vivo*, RoLFY est bien capable d'activer l'expression des gènes *RoAG* et *RoSHP*, médiateurs de la fonction C. Si cela est confirmé, la situation se révèlerait encore différente de celle d'*Arabidopsis* ou d'*Antirrhinum*, puisqu'ici les deux gènes *RoAG* et *RoSHP* participeraient à la fonction C de manière équivalente, et seraient tous les deux régulés par LFY. Il existerait donc un ensemble de choix possibles pour l'évolution de deux paralogues proches dans le contrôle d'une fonction aussi essentielle qu'est celle du développement des organes reproducteurs (Causier et al., 2005).

c) Liaison de LFY aux gènes B et C chez *Welwitschia mirabilis*, une gymnosperme

Une dernière étude, concernant la plante gymnosperme *Welwitschia mirabilis*, nous a permis de proposer des hypothèses sur l'apparition du réseau floral contrôlé par LFY chez les plantes terrestres (**Article 4**, p. 88). Les gymnospermes ne développent pas de fleurs, mais des cônes mâles ou femelles séparés. Comprendre le rôle de LFY dans la formation de ces

structures reproductrices peut nous donner des indices quant à l'apparition de la fleur et du réseau floral. Dans cette étude, nous avons pu isoler les gènes *WelB1*/*WelB2* et *WelC*, gènes MADS homologues respectivement aux gènes B et C des angiospermes, et nous avons analysé leurs domaines d'expression ainsi que ceux des gènes *WelLFY* et *WelNDLY*. Ces deux derniers gènes sont tous les deux exprimés dans les cônes mâles, et l'expression de *WelLFY* coïncide avec celle de *WelB1/WelB2*, alors que *WelNDLY* est exprimé dans un domaine similaire à celui de *WelC* (**Figure 3**). Ces résultats suggèrent une spécialisation de WelLFY et WelNDLY respectivement dans le contrôle de l'expression des gènes B et C. En effet, la spécificité de liaison à l'ADN des protéines WelLFY et WelNDLY apparaît différente en gel retard (**Figure 4D**), et WelLFY présente une spécificité (déterminée par les expériences de SELEX détaillées précédemment) similaire à celle de LFY d'*A. thaliana* (**Figure 5A**). Par la technique de Résonance Plasmonique de Surface (SPR) (Moyroud et al., 2009), la liaison de WelLFY aux promoteurs des gènes *WelB1* et *WelB2* a été validée (**Tableau 1**). Des sites de liaison de WelLFY ont été identifiés dans ces fragments grâce à la matrice de WelLFY issue du SELEX, et la liaison de la protéine à certains de ces sites a été confirmée par gel retard (**Figure 5D**). Ainsi, la liaison de WelLFY aux gènes *WelB1* et *WelB2* est validée par plusieurs approches *in vitro*, et est tout à fait cohérente avec les patrons d'expression des différents gènes, suggérant que WelLFY pourrait réguler l'expression des gènes *WelB1* et *WelB2* dans les cônes mâles. Ceci constitue les premiers éléments d'un réseau « pré-floral » impliquant LFY/NLY et les gènes B/C ; le contrôle des gènes B par LFY dans des tissus reproductifs était donc peut-être présent avant l'apparition de la fleur.

Cette étude est soumise pour publication au journal PNAS ; j'ai participé aux expériences de SELEX et de gels retards et à l'analyse des sites de liaison de WelLFY et WelNDLY dans les gènes B (**Figure 4**, **Figure 5**).

Le modèle que nous avons développé sur LFY-C peut donc s'appliquer à toutes les protéines LFY des plantes terrestres, si l'on connaît leur spécificité de liaison. Nous avons ainsi pu aboutir à des conclusions évolutives variées, chez des espèces non modèles, en utilisant uniquement la séquence génomique de quelques gènes.

3) Des prédictions à l'échelle génomique ?

a) Un modèle plus prédictif

Le modèle de calcul de scores que nous avons construit est hautement prédictif pour la liaison de LFY-C à l'ADN *in vitro*. Nous avons également constaté qu'à l'échelle génomique, jusqu'à 25% des sites prédits par notre modèle étaient effectivement liés par LFY *in vivo*. Ce degré de prédiction est de bonne qualité par rapport à ce qu'il existe dans la littérature, mais le nombre de faux positifs reste pourtant très élevé (75%). Il y a donc un saut important entre la prédiction des liaisons *in vitro* et *in vivo*, notamment car les nucléosomes, l'état de compaction de la chromatine, ou encore les autres facteurs de transcription vont moduler l'accessibilité de LFY à certains sites de liaison qui sont pourtant de bonne affinité pour la protéine *in vitro*.

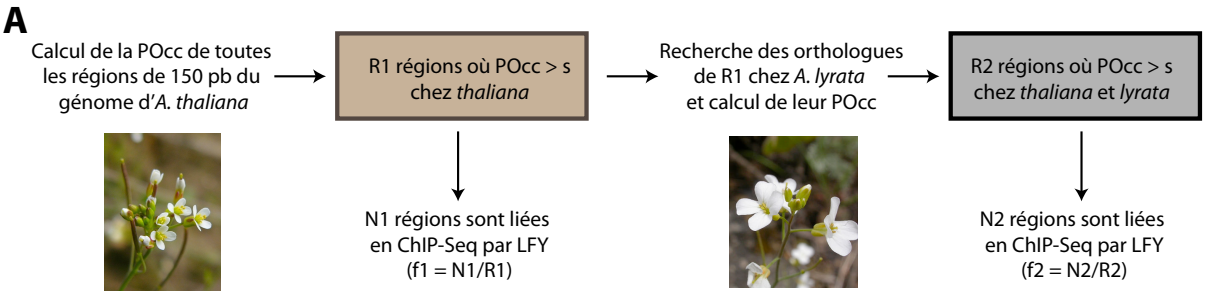
Comment savoir alors quels sites de liaison sont fonctionnels ? Comme discuté dans l'introduction, on imagine que les sites de liaison conservés au cours de l'évolution ont de grandes chances d'être fonctionnels; cette hypothèse est le fondement de la technique de phylogenetic footprinting (Gumucio et al., 1992; Hardison and Taylor, 2012). Pourtant, même si cette approche a fait ses preuves pour l'étude de gènes individuels (Lenhard et al., 2003), nous avons vu précédemment que la conservation de séquence seule n'était pas toujours suffisante pour détecter des sites de régulation, mais qu'il fallait d'abord pouvoir prédire la position des sites de liaison de façon fiable. J'ai donc utilisé le modèle bâti pour LFY-C pour détecter ses sites de liaison dans le génome d'*Arabidopsis thaliana*, et déterminer s'ils étaient conservés chez une espèce proche, *Arabidopsis lyrata*. En effet, pour pouvoir comparer la position de sites de liaison entre deux génomes, il faut que les relations d'orthologies entre les gènes (et même entre les régions régulatrices) soient claires, ce qui est le cas pour *A. thaliana* et *A. lyrata*. Afin de s'affranchir de la fluidité des sites de liaison que nous avons observée précédemment, j'ai à nouveau utilisé le calcul de l'Occupation Prédite (POcc) pour identifier les régions favorables à une liaison par LFY.

b) Conservation de la liaison de LFY aux génomes d'*A. thaliana* et *A. lyrata*

J'ai donc calculé la valeur de POcc de toutes les séquences glissantes de 150 pb du génome d'*A. thaliana*, et j'ai gardé celles où la POcc était supérieure à un seuil donné. J'ai ensuite cherché les orthologues de ces régions chez *A. lyrata*, et gardé à nouveau celles où la POcc était supérieure au seuil choisi. La même valeur de seuil de POcc peut être utilisée pour

les deux génomes, puisque leur POcc moyenne est très proche. En utilisant l'information de deux génomes, on espère isoler des régions plus susceptibles d'être liées par LFY *in vivo* ; j'ai donc recherché lesquelles des régions isolées, avec un ou deux génomes, étaient liées par LFY en ChIP-Seq (**Fig. 17A**).

A



B

s (seuil de POcc)	Nombre de régions R1	N1	f1 (%)	Nombre de régions orthologues à R1	Nombre de régions R2	N2	f2 (%)
0,04	93	27	29,03	77	10	4	40,00
0,02	479	86	17,95	366	39	12	30,77
0,01	2232	239	10,71	1714	190	43	22,63
0,005	11006	584	5,31	8092	1027	106	10,32
0,001	83256	1523	1,83	52835	30356	619	2,04

C

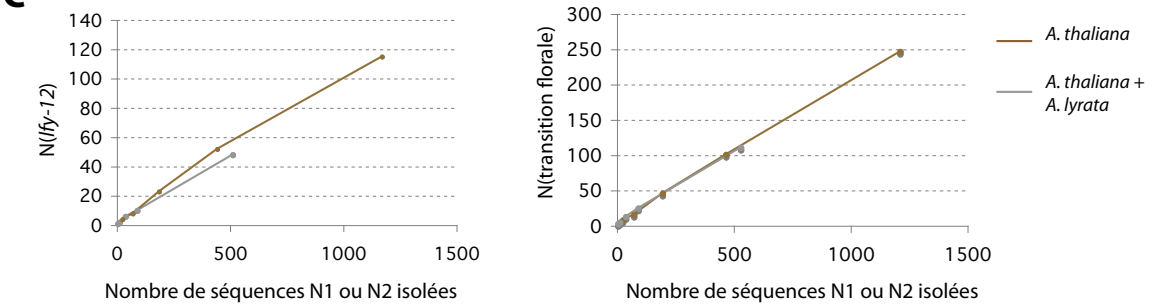


Figure 17 : Conservation de la liaison de LFY à l'échelle génomique entre *A. thaliana* et *A. lyrata*. A : Principe de la méthode employée. **B :** Nombre de régions isolées et liées en ChIP-Seq en fonction du seuil de POcc choisi, grâce au génome d'*A. thaliana* seul (marron) ou grâce aux deux génomes d'*A. thaliana* et d'*A. lyrata* (gris). **C :** Nombre de gènes dérégulés chez un mutant *lfy-12* (gauche, N(*lfy-12*)) ou lors de la transition florale (droite, N(transition florale)), en fonction du nombre de séquences isolées avec le génome d'*A. thaliana* seul (marron) ou avec les deux génomes d'*A. thaliana* et d'*A. lyrata* (gris). Le niveau de dérégulation N(*lfy-12*) d'un gène est le rapport d'expression de ce gène entre des plantes sauvages et des plantes *lfy-12* à 7 jours de développement, d'après des données publiques de microarray (Schmid et al., 2005). La valeur N(transition florale) d'un gène est le rapport des niveaux d'expression de ce gène chez des plantes à 0 jours ou 7 jours après un passage de Jours Courts à Cours Longs (changement de photopériode qui induit rapidement la floraison) (Schmid et al., 2003; Schmid et al., 2005). Un gène est dit dérégulé si N(*lfy-12*) est supérieur à 1,8 ou inférieur à 0,55, ou si N(transition florale) est supérieur à 2 ou inférieur à 0,5.

J'ai constaté que les régions détectées avec les deux génomes sont plus souvent liées en ChIP-Seq par LFY que celles isolées avec le génome d'*A. thaliana* seul. Par exemple, en utilisant un seuil de POcc de 0,01, 10,7% des régions isolées avec le génome d'*A. thaliana* sont liées en ChIP-Seq, alors que ce nombre atteint 22,6% en utilisant les deux génomes (**Fig.**

17B). Pourtant, le nombre de régions isolées est beaucoup plus faible puisqu'il est de 239 avec un génome, contre 43 avec deux génomes. Ainsi, l'utilisation des deux génomes va permettre de repérer des séquences peut-être plus fréquemment liées *in vivo*, mais en bien moins grand nombre. En réalité, ce comportement est le même lorsqu'on utilise un seul génome et qu'on augmente la valeur du seuil d'Occupation Prédite.

Les séquences détectées avec un génome et une haute valeur de POcc, ou avec deux génomes et une valeur de POcc plus faible, sont-elles les mêmes ? Si on isole un nombre comparable de séquences - par exemple 48 avec le génome d'*A. thaliana* seul avec un seuil de POcc à 0,027, et 47 en utilisant les deux génomes et un seuil à 0,01 - seulement 15 régions sont en fait présentes dans les deux jeux de séquences à la fois : les régions isolées avec un ou deux génomes ne sont donc pas identiques. Utiliser deux génomes devrait théoriquement permettre de repérer des régions dont la liaison a une importance fonctionnelle (puisque conservée). Pourtant, la proportion de gènes dérégulés chez un mutant *lfy-12* ou bien pendant la transition florale (Schmid et al., 2003; Schmid et al., 2005), qu'ils soient associés à des régions isolées avec un ou deux génomes, est similaire (**Fig. 17C**). La différence entre les régions à forte POcc dans un ou deux génomes n'a donc pas pu être mise en évidence pour le moment.

Ainsi, l'utilisation de deux génomes n'a pas ici permis d'améliorer le pouvoir prédictif de notre modèle et de détecter plus efficacement des régions liées ou régulées par LFY *in vivo*. Pourtant, l'utilisation de nombreux génomes pourrait très vraisemblablement décroître le taux de faux positifs détectés, et le nombre de génomes angiospermes séquencés augmente rapidement. La difficulté réside dans l'établissement des relations d'orthologie à l'échelle génomique, encore mal élucidées entre des espèces assez lointaines.

Les séquences issues du SELEX réalisé sur les différentes protéines LFY nous ont permis de construire un modèle hautement prédictif pour la liaison de LFY in vitro*. Ce modèle pourrait encore être optimisé pour obtenir une meilleure prédiction des cibles de LFY* in vivo *; j'ai tenté de l'améliorer sans succès, mais de nombreuses autres possibilités restent à tester. Nous avons utilisé ce modèle pour étudier l'évolution de la régulation de quelques gènes par LFY chez les angiospermes et les gymnospermes, ce qui nous a permis d'aboutir à des conclusions biologiques et évolutives variées, en utilisant uniquement des séquences génomiques. Nous avons également observé à plusieurs reprises que les éléments* cis *de liaison de LFY étaient peu conservés en séquence et en score même chez des espèces très*

proches, soulignant une grande fluidité au niveau de ces éléments alors même que le réseau n'a pas évolué en trans.

Articles complémentaires à ce chapitre

Article 2 : Prediction of Regulatory Interactions from Genome Sequences Using a Biosphysical Model for the *Arabidopsis* LEAFY Transcription Factor
Edwige Moyroud, Eugenio Gómez-Minguet, Felix Ott, Levi Yant, David Posé, Marie Monniaux, Sandrine Blanchet, Olivier Bastien, Emmanuel Thévenon, Detlef Weigel, Markus Schmid et François Parcy.
The Plant Cell, Vol. 23:1293-1306

Article 3 : LEAFY Target Genes Reveal Floral Regulatory Logic, *cis* Motifs, and a Link to Biotic Stimulus Response
Cara Winter, Ryan Austin, Servane Blanvillain-Baufumé, Maxwell Reback, Marie Monniaux, Miin-Feng Wu, Yi Sang, Ayako Yamaguchi, Nobutoshi Yamaguchi, Jane Parker, François Parcy, Shane Jensen, Hongzhe Li et Doris Wagner.
Developmental Cell, Vol. 20:430-443

Article 4 : A link between LEAFY and B genes in *Welwitschia mirabilis* sheds light on ancestral mechanisms prefiguring floral development
Edwige Moyroud, Marie Monniaux, Emmanuel Thévenon, Renaud Dumas, Charles Scutt, Michael Frohlich et François Parcy.
Soumis à PNAS le 26/09/12

The Plant Cell, Vol. 23: 1293–1306, April 2011, www.plantcell.org

Prediction of Regulatory Interactions from Genome Sequences Using a Biophysical Model for the *Arabidopsis* LEAFY Transcription Factor[C][W]

Edwige Moyroud,[a] Eugenio Gómez Minguet,[a,1] Felix Ott,[b] Levi Yant,[b,2] David Posé,[b] Marie Monniaux,[a] Sandrine Blanchet,[a] Olivier Bastien,[a] Emmanuel Thévenon,[a] Detlef Weigel,[b] Markus Schmid,[b] and François Parcy[a,3]

[a] Laboratoire de Physiologie Cellulaire Végétale, Unité Mixte de Recherche 5168, Centre National de la Recherche Scientifique, Commissariat à l'Énergie Atomique, Institut National de la Recherche Agronomique, Université Joseph Fourier Grenoble I, 38054 Grenoble, France

[b] Max Planck Institute for Developmental Biology, Department of Molecular Biology, 72076 Tuebingen, Germany

Despite great advances in sequencing technologies, generating functional information for nonmodel organisms remains a challenge. One solution lies in an improved ability to predict genetic circuits based on primary DNA sequence in combination with detailed knowledge of regulatory proteins that have been characterized in model species. Here, we focus on the LEAFY (LFY) transcription factor, a conserved master regulator of floral development. Starting with biochemical and structural information, we built a biophysical model describing LFY DNA binding specificity in vitro that accurately predicts in vivo LFY binding sites in the *Arabidopsis thaliana* genome. Applying the model to other plant species, we could follow the evolution of the regulatory relationship between LFY and the *AGAMOUS* (*AG*) subfamily of MADS box genes and show that this link predates the divergence between monocots and eudicots. Remarkably, our model succeeds in detecting the connection between LFY and *AG* homologs despite extensive variation in binding sites. This demonstrates that the *cis*-element fluidity recently observed in animals also exists in plants, but the challenges it poses can be overcome with predictions grounded in a biophysical model. Therefore, our work opens new avenues to deduce the structure of regulatory networks from mere inspection of genomic sequences.

INTRODUCTION

New technologies rapidly deliver whole-genome sequences from a wide variety of organisms at low cost, but functional annotation of these genomes remains a major challenge. Whereas conserved protein sequences are easily identified, transcriptional *cis*-regulatory modules can be evolutionarily fluid (Wilson and Odom, 2009; Schmidt et al., 2010; Weirauch and Hughes, 2010). Several recent studies revealed significant divergence in binding profiles of transcription factor (TF) homologs between vertebrate species (Mikkelsen et al., 2010; Schmidt et al., 2010). This divergence is due to the nature of *cis*-elements, which are small and degenerate motifs that can change rapidly and are thus difficult to detect by simple DNA sequence comparison (Wasserman and Sandelin, 2004; Ward and Bussemaker, 2008; Badis et al., 2009; Wilson and Odom, 2009). Whereas it is possible to study the genome-wide binding profile of TFs to DNA experimentally using chromatin immunoprecipitation (ChIP), a more streamlined functional analysis of genomes requires methods to predict variable *cis*-elements accurately directly from DNA sequences.

To address this problem, we focused on the genetic circuitry downstream of the LEAFY (LFY), a TF with a central role in the evolution and development of flowers (Liu et al., 2009; Moyroud et al., 2010). In *Arabidopsis thaliana*, LFY directly activates the expression of several floral homeotic MADS box genes, including *AGAMOUS* (*AG*), *APETALA1* (*AP1*), and *AP3* (Parcy et al., 1998; Busch et al., 1999; Wagner et al., 1999; Lohmann et al., 2001; Lamb et al., 2002), while repressing the shoot program by downregulating genes such as *TERMINAL FLOWER1* (*TFL1*) (Liljegren et al., 1999; Ratcliffe et al., 1999; Parcy et al., 2002). From the small number of known LFY DNA binding sites, only a poorly defined 7-bp consensus sequence, CCANTG[G/T], has been previously deduced (Busch et al., 1999; Lamb et al., 2002). The three-dimensional structure of the LFY DNA binding domain has revealed contacts over 19 bp, suggesting considerably greater specificity (Hamès et al., 2008). Our aim was to capture this specificity in a predictive tool capable of detecting LFY binding sites from plant genomic sequences and ultimately tackle evolutionary questions. Here, we show how a biophysical model, built on biochemical ground and optimized using

[1] Current address: Instituto de Biologia Molecular y Celular de Plantas (UPV-CSIC), Universidad Politécnica de Valencia, Avda de los Naranjos s/n, Valencia 46022, Spain.

[2] Current address: Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138.

[3] Address correspondence to francois.parcy@cea.fr.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: François Parcy (francois.parcy@cea.fr).

[C] Some figures in this article are displayed in color online but in black and white in the print edition.

[W] Online version contains Web-only data.

www.plantcell.org/cgi/doi/10.1105/tpc.111.083329

genome-wide in vivo binding data, can predict the evolution of the relationship between LFY and *AG* homologs, despite extensive variation in the sequences and positions of binding sites.

RESULTS

A Model for LEAFY DNA Binding Specificity

We determined the DNA binding preferences of the LFY DNA binding domain (DBD) using high-throughput systematic evolution of ligands by exponential enrichment (Selex) (Figure 1A) (Zhao et al., 2009). Alignment of the 494 unique sequences obtained revealed a 19-bp motif (Figure 1C), in good agreement with the three-dimensional structure of LFY DBD complexed with DNA (Hamès et al., 2008). This motif displays the previously established 7-bp consensus as the core. From the alignment, we deduced an asymmetric (ASY) position-specific scoring matrix (PSSM) (Wasserman and Sandelin, 2004) (Figure 1C; see Supplemental Table 1 online). Using this matrix with any 19-bp DNA fragment, scores can be calculated that should be proportional to the logarithm of the affinity of LFY DBD for this fragment. We used quantitative multifluorescence relative affinity (QuMFRA) assays (Man and Stormo, 2001) to measure the relative affinity of LFY DBD for 48 different oligonucleotides. We found that the ASY matrix scores correlated well with experimentally measured DNA binding affinities (Pearson correlation, $r^2 = 0.59$) (Figure 1C). Since the LFY DBD binds DNA as a symmetric homodimer (Hamès et al., 2008), we sought to improve the PSSM by imposing symmetry. With the corresponding SYM matrix (Figure 1D), r^2 increased to 0.69. To improve the matrix predictive power further, we analyzed the dependence between nucleotide positions: simple PSSMs assume that different positions contribute independently to the overall binding, but this condition is not always satisfied (Benos et al., 2002). For LFY, we indeed observed nonindependent triplets at two symmetric positions and in the center of the alignment (Figure 2). We modeled this dependence using the frequency of trinucleotides (Figure 1E).

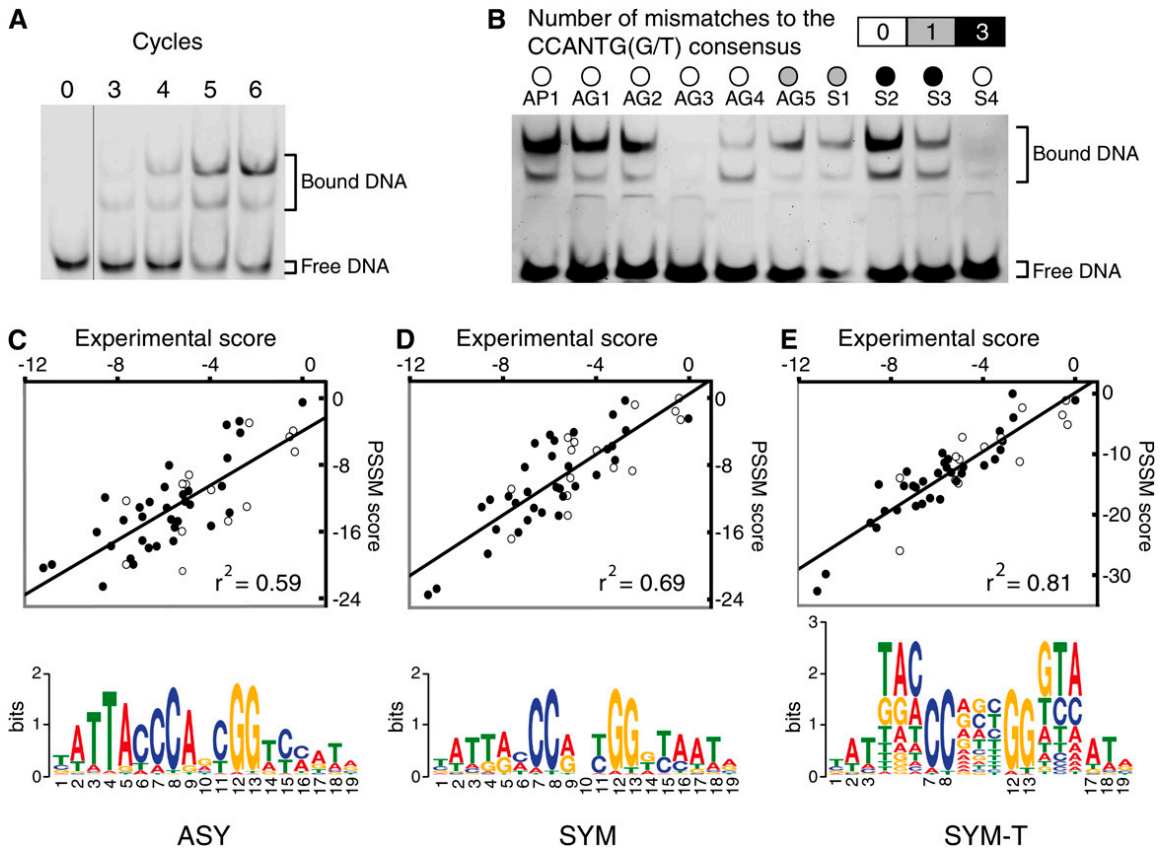


Figure 1. Optimization of the LFY Binding Site Model.

(A) Enrichment of DNA sequences bound by LFY over different Selex cycles.

(B) Binding of LFY to different sequences, either from *AG* or *AP1* genes, or synthetic (S), with varying numbers of mismatches to the previously recognized consensus LFY binding motif.

(C) to **(E)** Comparison of experimentally determined and predicted scores (see Methods) for different DNA sequences with the three PSSMs (asymmetric [ASY], symmetric [SYM], and symmetric with triplets [SYM-T]), illustrated below by their logos. Open and closed circles represent sequences with or without the CCANTG[G/T] consensus, respectively.

[See online article for color version of this figure.]

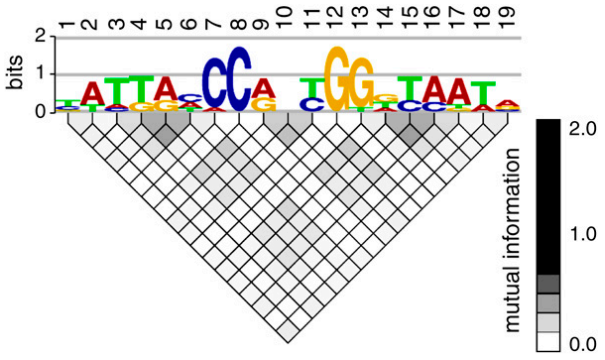


Figure 2. Detection of Dependence between Positions of the LFY Binding Sites.

Alignment of the 494 Selex sequences was analyzed with enoLOGOS software (Workman et al., 2005). The mutual information of each pair of positions of the alignment is displayed as a gray-scale-coded matrix plot below the logo corresponding to the SYM PSSM. Dependence is detected between positions 4, 5, and 6 or 14, 15, and 16 (lateral triplets) and, to a lesser extent, between positions 9, 10, and 11 (central triplet). [See online article for color version of this figure.]

The resulting SYM-T matrix further increased r^2 to 0.81. Notably, whereas the SYM-T matrix was well correlated with experimental DNA binding affinities, the simple presence or absence of the 7-bp consensus motif in the oligonucleotides tested was a poor predictor of binding, confirming the usefulness of the PSSM approach (Figures 1B to 1E).

In Vivo Validation of the LFY Model by ChIP-seq

To test how well the in vitro–determined DNA binding specificity correlated with in vivo binding, we performed a ChIP experiment with LFY-specific antibodies followed by short read sequencing (ChIP-seq). The genomic regions enriched in plants that overexpressed LFY (*35S:LFY*) compared with wild-type seedlings were ordered using the rank product from two ChIP-seq replicates. In parallel, we used a biophysical model to compute the predicted occupancy (POcc) of these genomic regions by LFY (Granek and Clarke, 2005; Ward and Bussemaker, 2008). Such a model uses a PSSM to estimate the scores of all binding sites present on a large DNA fragment and then integrates these scores to compute the POcc value. The regions identified in ChIP-seq were ranked according to their POcc. We found a good correlation between the prediction and the experimental ChIP-based ranking. Moreover, we observed that the correlation increased from the ASY (Spearman's rank correlation coefficient, 0.44) and the SYM (0.45) to the SYM-T matrix (0.53).

As further validation, we performed a receiver operating characteristic (ROC) analysis (Hanley and McNeil, 1982) comparing the 1564 regions most strongly enriched in ChIP (false discovery rate [FDR] < 0.1 in each of two independent replicates, meaning that the FDR is lower than 0.01 on the whole experiment for each gene selected; see Supplemental Data Set 1A online) with a set of random nonbound negative regions. In this analysis, we compared the percentage of regions whose POcc is higher than a given threshold in bound and unbound fragments sets. The area under the curve (ROC AUC) quantifies the tradeoff of specificity and sensitivity of the model as the POcc threshold varies. We evaluated the performance of two versions of the biophysical model: a first one that integrates all sites present on the fragment and a second one (hit-based model) that selects binding sites with a score higher than a cutoff value (Roider et al., 2007). With a ROC AUC value of 0.865 (Figure 3), the second model was best, but both of our models performed very well compared with other studies where ROC AUC values higher than 0.85 are found for <15% of the TFs studied (Granek and Clarke, 2005; Roider et al., 2007).

LFY Directly Binds to Key Genes Regulating Flower Development

The most highly ranked ChIP-enriched fragment was in the 3′ region of the *TFL1* gene, which is repressed by LFY and has important regulatory elements downstream of the transcribed region (Ratcliffe et al., 1999; Kaufmann et al., 2010). The strong binding observed in ChIP is explained by the presence of a cluster of LFY binding sites missing the CCANTG[G/T] consensus but detected by the SYM-T model (Figure 4B). Another very highly ranked region was present in the promoter of the well-characterized target *AP1* (Parcy et al., 1998; Wagner et al., 1999), which also showed a second peak due to the presence of a binding site in its first intron (Figure 4A). These two results strongly suggest that LFY represses *TFL1* both directly, as proposed before based on experiments with an activated form of LFY (Parcy et al., 2002), and indirectly, through *AP1* activation (Kaufmann et al., 2010). For both *AP1* and *TFL1* as for most of the regions examined, the similarity between the ChIP-seq profiles and the computed binding site landscapes was striking (Figure 4), underscoring the predictive power of the SYM-T binding model.

The ChIP experiment also identified binding of LFY to regulatory regions of numerous floral regulator genes, such as *AG* (Busch et al., 1999) and *SEPALLATA4* (Figures 4C and 4D) but also *LFY* itself (suggesting autoregulation) and *GLABROUS*

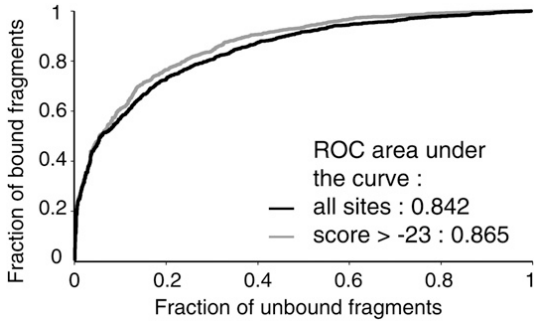


Figure 3. Comparison of the Different Models for Prediction of in Vivo LFY Binding Sites.

ROC curves for LFY-bound and unbound sequences, using a biophysical model taking all sites (black line) into account or only those with a SYM-T matrix score higher than −23 (gray line).

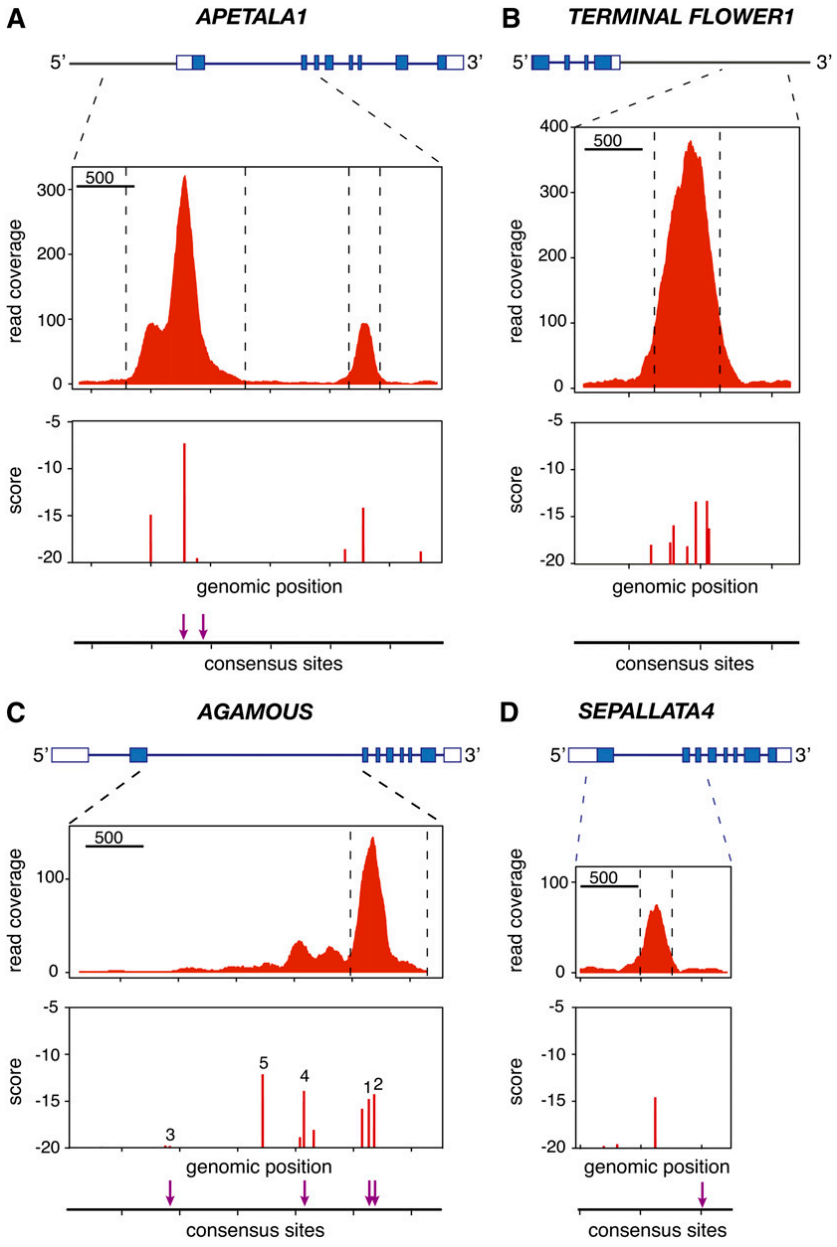


Figure 4. Examples of LFY-Bound Regions Identified by ChIP-seq.

Noncoding and coding sequences in exons are shown on top as open and closed boxes, respectively. ChIP-seq read coverage combined from both strands is shown in the middle. The bottom panels show the scores of binding sites (computed with the SYM-T model) and the presence of the CCANTG[G/T] consensus (indicated by arrows). *AP1* **(A)**, *TFL1* **(B)**, *AG* **(C)**, and *SEP4* **(D)**.
[See online article for color version of this figure.]

INFLORESCENCE STEMS (Gan et al., 2006) (see Supplemental Figure 1 online; Table 1). Bound regions were also found in genes related to gibberellins and auxin signaling, two hormones known to be important for flower development (see Supplemental Figure 1 online; Table 1). Among the 2677 genes adjacent to the 1564 bound regions (see Supplemental Data Sets 1B and 1C online), 320 genes have an altered expression in *lfy* mutants (Schmid et al., 2003) and 54 (out of 445 genes; P value = 0.025) are deregulated in *LFY-GR*–overexpressing plants (William et al., 2004) (see Supplemental Data Set 1C online), including nine of the 15 genes previously considered as LFY direct targets by William et al. (2004). We expect many of the genes that are both bound and regulated to represent bona fide LFY direct target genes. In most cases, our model identified the LFY binding sites potentially responsible for the signal observed in ChIP (Figure 4; see Supplemental Figure 1 online).

Table 1. Examples of Genes Bound by LFY

Gene	Primary Gene Symbol	Rank	POcc	Best Site
Flowering				
At5G03840	*TFL1*	1	0.01803	−13.60
At2G45660	*SOC1*	770	0.0098	−12.49
At2G39250	*SCHNARCHZAPFEN*	1029	0.0070	−13.95
At3G58070	*GLABROUS INFLORESCENCE STEMS*	259	0.0034	−19.26
		288	0.0225	−10.25
		1482	0.0024	−19.25
At1G01183	*miR156*	8	0.0085	−13.61
At4G35900	*FD*	796	0.0037	−15.69
At4G01500	*NGATHA4*	725	0.0081	−13.49
At1G25560	*TEM1*	6	0.0048	−17.12
		498	0.0026	−19.40
At4G25520	*SLK1*	68	0.0042	−16.63
At2G45190	*FILAMENTOUS FLOWER*	503	0.0030	−16.45
Floral meristem specification				
At5G61850	*LFY*	1269	0.0156	−10.82
At3G57130	*BOP1*	165	0.0476	−7.82
		1400	0.0026	−18.83
At2G41370	*BOP2*	166	0.0051	−16.05
		556	0.0034	−20.12
		671	0.0112	−11.83
At5G18560	*PUCHI*	574	0.0046	−18.27
Floral organ specification and development				
At1G69120	*APETALA1*	19	0.0609	−7.33
		1216	0.055	−14.22
At4G18960	*AG*	888	0.0104	−14.28
At3G54320	*WRINKLED1*	815	0.0041	−14.95
At1G24260	*SEPALLATA3*	25	0.0096	−16.15
		829	0.0023	−20.59
At2G03710	*SEPALLATA4*	983	0.0049	−14.70
At1G31140	*GORDITA*	1421	0.0403	−8.20
At5G02030	*PENNYWISE*	1221	0.0043	−14.77
At3G63530	*BIG BROTHER*	940	0.0110	−12.23
		989	0.0016	−19.54
At5G67060	*HECATE1*	527	0.0068	−13.97
At4G36260	*STYLISH 2*	520	0.0071	−13.81
At5G07280	*EMS1*	772	0.0044	−14.70
At3G02000	*ROXY1*	367	0.0084	−14.60
At2G28056	*miR172*	1213	0.0054	−16.79
At2G28610	*PRESSED FLOWER*	424	0.0166	−12.23
At4G37750	*AINTEGUMENTA*	462	0.0024	−20.34
		1460	0.0034	−18.39
At5G10510	*AINTEGUMENTA-like 6*	1653	0.0056	−14.89
At1G01510	*ANGUSTIFOLIA 3*	723	0.0019	−21.45
Gibberellins				
At5G15230	*GASA4*	231	0.0095	−12.69
		431	0.0062	−13.83
At4G25420	*GA2OX1*	1427	0.0086	−13.52
At1G30040	*GA2OX2*	1045	0.0074	−13.94
At3G63010	*GID1B*	350	0.0068	−14.69
		425	0.0052	−14.82
		1536	0.0025	−17.42
At1G15550	*GA3OX1* (*GA4*)	1573	0.0056	−17.72
At1G80340	*GA3OX2*	879	0.0052	−15.90
Auxin				
At1G19840	*SAUR-like*	263	0.0416	−8.21

(Continued)

Table 1. (continued).

Gene	Primary Gene Symbol	Rank	POcc	Best Site
At1G19850	*MONOPTEROS*	289	0.0217	−10.15
At2G01420	*PIN4*	235	0.0029	−17.34
		999	0.0053	−15.03
At3G62980	*TIR1*	100	0.0107	−12.78
		110	0.0033	−17.03
At5G11320	*YUCCA4*	510	0.0061	−15.82
		1261	0.0055	−14.73
At1G04240	*SHY2*	1350	0.0043	−15.32
At2G34650	*PINOID*	225	0.0040	−15.83
At1G29430	Auxin-responsive (*SAUR-like*)	212	0.0228	−9.85
		438	0.0151	−11.53
Cytokinins				
AT1G59940	*ARR3*	1088	0.0204	−10.13

For a selection of genes expressed in floral tissues or dependent on LFY, the table indicates the rank from the ChIP-seq experiments (Rank), the POcc value, and the score of the best LFY binding site. Binding profiles are shown in Figure 4 or Supplemental Figure 1 online for the genes with underlined names.

Analysis of the LEAFY-*AG* Link over Large Evolutionary Distances

A major motivation for developing predictive DNA binding models is the functional annotation of genomes from nonmodel organisms. For a proof of concept, we examined the large intron of *AG* homologs, since this region is known to be important for *AG* regulation in various species and contains several conserved motifs (Sieburth and Meyerowitz, 1997; Busch et al., 1999; Davies et al., 1999; Hong et al., 2003; Causier et al., 2008). *AG* belongs to a small subfamily of MADS box genes (Ferrario et al., 2004; Zahn et al., 2006). A first duplication led to the formation of the *AG* and *AGL11* lineages at the base of the angiosperms, and a second duplication in ancestral core eudicots yielded the *euAGAMOUS* (*euAG*) and *PLENA* (*PLE*) lineages (Kramer et al., 2004) (Figure 5A). All these proteins have similar DNA binding and protein–protein interaction profiles, and it is thought that they evolved specific functions primarily through diversification of their expression patterns (Ferrario et al., 2004; Zahn et al., 2006). Sequence similarity and genomic position are therefore not sufficient to predict functional equivalence with *AG* in other species.

As the structural models indicated that the LFY-DNA interface is highly conserved in angiosperms (Moyroud et al., 2009), we applied our threshold-based biophysical model to the large intron of *AG* subfamily members of multiple angiosperm species. In both *A. thaliana* and its relative *Arabidopsis lyrata*, the predicted occupancy by LFY is much higher for the *AG* second intron than for that of *SHATTERPROOF* (*SHP1* and *SHP2*, belonging to the *PLE* lineage) and *SEEDSTICK* (*STK*; belonging to the *AGL11* lineage) genes (Figure 5C). This prediction is validated by functional analyses in *A. thaliana* demonstrating that LFY is responsible for the early induction of the *AG* gene (Parcy et al., 1998; Busch et al., 1999; Lohmann et al., 2001) but is not involved in regulating *SHP* or *STK* genes, which play later roles in fruit and ovule development (Liljegren et al., 2000; Colombo et al., 2010). Consistent with this, only *AG*, but not *SHP* or *STK*, was found to be a LFY target in our ChIP-seq experiments. Conversely, in several eudicots, such as *Antirrhinum majus* or *Solanum lycopersicum*, genes from the *PLE* clade were found to have the highest POcc compared with *euAG* or *STK* genes (Figure 5C). Our analysis thus predicts that they should be regulated by LFY. This prediction has indeed been validated in *A. majus*, where the *SHP* ortholog *PLE* was shown to be activated by the LFY ortholog FLORICAULA and to have an *AG*-like function (Davies et al., 1999; Causier et al., 2005). In other eudicot species, where less functional data is available, we observed a good agreement between a high POcc by LFY and the expression of the corresponding genes during early stages of flower development, when *LFY* is active (Figure 5C; see Supplemental Table 2 online).

We also examined *AG* and *AGL11* orthologs from grasses, which are monocots. In all species examined, our model predicts much higher DNA occupancy by LFY for both *AG* orthologs compared with those of *AGL11* (Figure 5B). This prediction is validated by expression data and functional analyses demonstrating that, in grasses, *AG* genes are both expressed before *AGL11* orthologs and share the C-function (see Supplemental Table 2 online) (Thompson and Hake, 2009). Also, genetic analyses have suggested that *ZFL1/2*, the *LFY* maize (*Zea mays*) orthologs, regulate *AG* genes expression (Bomblies et al., 2003).

Detection of *cis*-Element Fluidity in *AG* Introns

Whereas our model correctly predicts global LFY occupancy in the large introns of *AG* homologs, we observed that the binding site landscapes are highly variable between these genes (Figure 6; see Supplemental Figures 2 to 4 online). In some cases, such as Bd-*AG* and Vv-*AG2*, there is a single binding site of very high affinity (corresponding to the AG2 LFY binding site in *A. thaliana*; Busch et al., 1999), whereas in others, such as At-*AG*, Al-*AG*, Os-*MADS58*, or *PMADS3*, this site is present but has a lower affinity that is compensated for through the action of multiple other sites (Figure 6). We experimentally verified the predicted high affinity

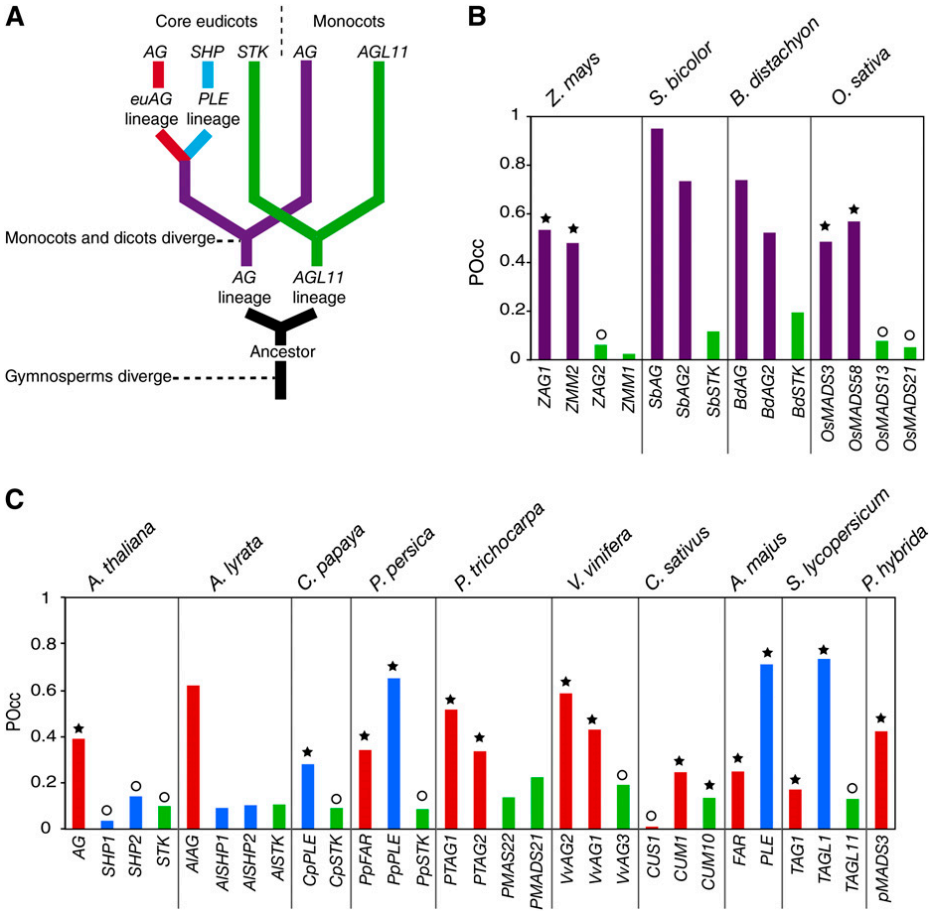


Figure 5. Prediction of LFY Occupancy of the Large Intron of *AG* Homologs Using the SYM-T Model.

(A) Schematic phylogeny of *AG* homologs after Kramer et al. (2004).
(B) and **(C)** POcc of *AG* homologs in monocots **(B)** and eudicots **(C)**. A star indicates gene expression during early floral stages, and a circle indicates later expression. Expression data come from the references listed in Supplemental Table 2 online.

for LFY for some of these additional binding sites (AG1, AG4, and AG5 from *A. thaliana AG*) (Figure 1B). We also detected their presence in multiple Brassicaceae species (see Supplemental Figure 4 online), strongly suggesting that they are functionally relevant.

Next, we aligned the introns of *AG* homologs using the DIALIGN program (Morgenstern, 2004), which allows identification of local sequence similarities in divergent sequences. The highest-affinity binding site (corresponding to AG2 in *A. thaliana*) can be detected in alignments, but the sequence conservation is fairly low with many more regions of higher conservation spread throughout the intron (see Supplemental Figures 2 to 4 online). The other LFY binding sites cannot be identified based on sequence conservation alone, even in plants belonging to the same family such as the Brassicaceae (see Supplemental Figures 2 to 4 online). These results illustrate the fluidity of binding sites and the difficulty of detecting them by sequence alignment, in agreement with recent comparative genome-wide analyses of TF binding sites in vertebrates (Mikkelsen et al., 2010; Schmidt et al., 2010). The strength of a biophysical model is to overcome *cis*-element plasticity and detect regulatory links despite extensive sequence variation.

DISCUSSION

In this work, we built a model for DNA recognition by the LFY TF. The core tools we used (PSSMs and biophysical models) were developed and validated for bacterial and animal TFs (Wasserman and Sandelin, 2004) and have rarely been used in plant studies. The originality of our work resides in the fact that we have incorporated structural information (to impose the PSSM symmetry) and the dependence between nucleotides, thereby generating an improved model with high predictive power both for in vitro and in vivo binding (Figures 1 to 3). The fact that the PSSM built in vitro using LFY DBD explains very well the ChIP-seq results obtained with the full-length LFY protein strongly suggests that LFY DBD contains most of the DNA binding specificity.

Among the various methods available to build PSSMs, reiterative in vitro selection of binding sites followed by PCR (Selex) is particularly well suited: for TFs with large binding sites such as

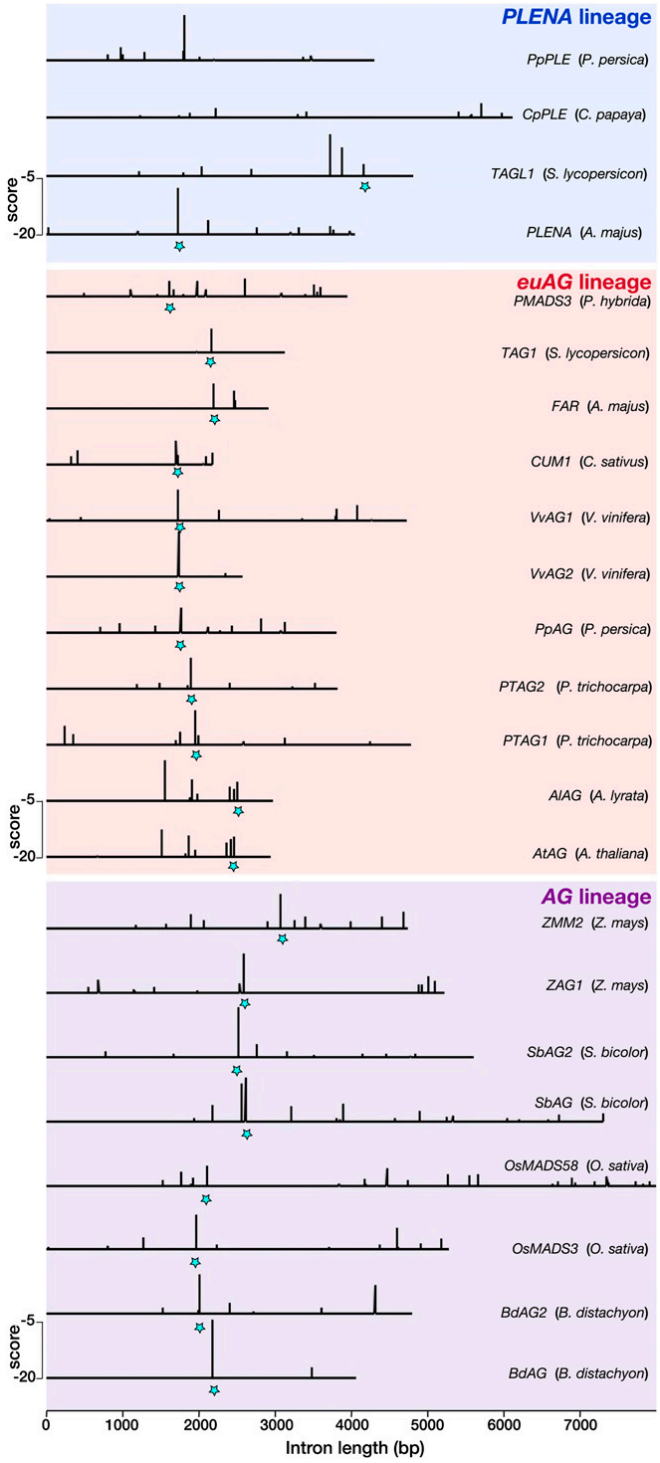


Figure 6. Distribution of LFY Binding Sites in *AG*-Like Genes.

LFY binding sites with a score higher than −20 are shown in eudicots (*PLENA* and *euAG* lineages) and monocots (*AG* lineage). The score scale is shown in each panel; the best binding sites correspond to the less negative score values. Stars mark the LFY binding site AG2, which can be located with confidence in most introns thanks to a nearby conserved sequence (see Supplemental Figure 2 online). Gene and species names are indicated on the right.

[See online article for color version of this figure.]

LFY, it is superior to the use of defined microarrays (Badis et al., 2009), which are limited in their complexity and cannot be reasonably used for binding sites larger than 11 nucleotides. Also, Selex allows the capture of important specificities that are not detected using ChIP experiments, such as the dependence between nucleotides. As illustrated in this and other studies (Figures 1 and 4; see Supplemental Figure 1 online), PSSMs (derived from Selex or ChIP experiments) are far superior to consensus sequences, which show poor predictive power and provide only binary information that cannot be incorporated into biophysical models.

To validate the in vitro–generated model, we performed a ChIP-seq experiment on seedlings constitutively expressing LFY. This experiment confirmed the quality of our model but is not sufficient to establish that all identified bound regions indeed correspond to genuine target genes. Still, many expected candidates, such as *AP1*, *AG*, or *TFL1*, have been identified with high confidence and the expression of several genes with bound regions changes in *lfy* mutants or plants overexpressing the LFY-GR inducible LFY protein (Wagner et al., 1999; Schmid et al., 2003). Combining the ChIP experiment with the biophysical model predictions allowed us to identify numerous previously unknown LFY binding sites that cannot be detected with the 7-bp consensus sequence (Figure 4; see Supplemental Figure 1 online). The good agreement observed in many cases between the location of these sites and the ChIP-seq peaks illustrates the capacity of our model to position the LFY binding sites correctly in genomic DNA sequence. Some cases remain where the model does not easily explain the in vivo LFY binding, suggesting that LFY might possess other modes of DNA binding (through contacts with another TF, for example).

We also used the LFY binding model to search the whole *A. thaliana* genome for high scoring binding sites or for regions with a high POcc (see Supplemental Table 3 online). Among the 100 highest-scoring sites in the genome, ~25% were found to be bound in ChIP-seq, and it is likely that this percentage would increase if the ChIP-seq experiments were performed with inflorescence tissues. This result further corroborates the unique performance of this model when applied to the whole genome. Lowering the score or the POcc threshold identifies numerous regions that were not bound in the ChIP-seq experiments (see Supplemental Table 3 online). A major cause for this discrepancy is probably the accessibility of DNA. As shown in other systems, the incorporation of DNA accessibility estimated from chromatin marks or nucleosome positioning is likely to improve the prediction of bound sites further (Whitington et al., 2009; Won et al., 2010).

The results we obtained in vitro and in *A. thaliana* plants demonstrate that our model is highly predictive and can be used to address evolutionary questions. We analyzed the relationship between LFY and one of its target genes (*AG*) in various species. We showed that the computation of the predicted occupancy (POcc), which integrates the influence of numerous binding sites over a large DNA region, enables us to predict the relationship between LFY and members of the *AG* subfamily solely based on genomic sequence analysis. The case of the grasses is particularly striking: in all species examined, the two *AG* paralogs show much higher POcc values than the *AGL11* genes do (Figure 5).

Based on the presence of one LFY consensus site in a single rice (*Oryza sativa*) *AG* paralog (Causier et al., 2008), it had been previously proposed that the regulation of *AG* by LFY could predate the divergence between monocots and dicots. We now confirm this hypothesis based on the analysis of eight *AG* genes from monocots. The power of the POcc computation is also illustrated in angiosperms: for all *AG*-like genes, we found a good agreement between expression during early flower meristem development (when LFY is active) and high POcc of the *AG* large intron by LFY. Our analysis could even differentiate between the functional homologs of *A. thaliana AG* in species such as *A. majus* or *S. lycopersicum* where a functional shift has occurred so that the *SHP* orthologs (*PLE* and *TAGL1*, respectively) participate in *AG*-like function.

In addition to the global analysis based on POcc computation, the examination of the distribution of individual LFY binding sites in *AG* introns also yielded interesting insights. In the Brassicaceae, the family to which *A. thaliana* belongs, a previous study analyzed the *AG* large second intron by phylogenetic shadowing, identifying several conserved regions (Hong et al., 2003). One of these regions included a conserved site (AG3; see Supplemental Figure 4 online) that exhibited the 7-bp consensus sequence CCANTG[G/T] and was therefore proposed to be a LFY binding site. We have now shown that it is not a bona fide LFY binding site (Figure 1). Conversely, our LFY PSSM identified a previously unrecognized site (AG5), for which we confirmed a high affinity of LFY in vitro (Figure 1). Neither this site nor the previously identified AG4 site (Hong et al., 2003) was bound in our ChIP experiment in seedlings, presumably because of their closed chromatin conformation: analysis of the H3K27 trimethylation repressive marks indeed has shown that in *A. thaliana* seedlings, only a short region encompassing the AG1 and AG2 sites is in open configuration (Zhang et al., 2007). Still, the presence in most Brassicaceae examined of the AG5 high-affinity site (with little sequence conservation of the site itself) (see Supplemental Figure 4 online), together with AG4 analysis in *A. thaliana* (Hong et al., 2003), strongly support their functional importance.

Comparing more distant species (Figure 6; see Supplemental Figures 2 and 3 online) revealed that the LFY/*AG* transcriptional link was conserved despite extensive variation in number, position, sequence, and affinity of individual binding sites. Several recent studies in animals have observed considerable variation in TF binding profiles between species. However, these differences do not seem to be systematically associated with changes in target gene expression (Odom et al., 2007; Wilson and Odom, 2009; Dowell, 2010; Kasowski et al., 2010; Weirauch and Hughes, 2010). A recent study examining TF binding in vertebrate genomes showed that conserved regulatory interactions do not increase sequence constraints (Schmidt et al., 2010). Therefore, *cis*-elements must be fluid; they can vary without necessarily compromising transcriptional regulation. This property represents an obstacle for approaches based on sequence conservation, such as genomic shadowing or phylogenetic footprinting (Wasserman and Sandelin, 2004). Our study shows that this fluidity also exists in plants but can be overcome using an integrative biophysical model, which detects regulatory interactions despite extensive *cis*-element plasticity.

As more plant genome sequences become available, it is essential to be able to derive functional information from direct examination of primary sequences. Our work illustrates the potential of biophysical models to predict regulatory interactions. Thanks to its relatively large binding site with high information content, LFY presents key advantages to pioneer such an approach. Nevertheless, it should be possible to generalize this type of analysis to other TFs provided that the PSSM have been established: biophysical models can easily incorporate cooperativity and competition between TFs and can be efficiently applied to combinations of TFs with smaller individual binding sites (Granek and Clarke, 2005). The case of heterodimeric TFs, such as MADS box factors, is obviously more complex: PSSMs could be derived from Selex procedures adapted to heterodimeric complexes or from ChIP experiments, but in the latter case, they would represent a mixture of the different complexes present in the tissue. Once successfully generalized to various types of TF, our strategy represents a powerful approach for both the functional annotation of genomes of nonmodel species and the prediction of regulatory network evolution directly from primary DNA sequences. It can be efficiently coupled to genome-wide expression data or comparison between species (Ward and Bussemaker, 2008; Yeo et al., 2009). In particular, it will be interesting to analyze genomic sequences from basal angiosperms, once available, to understand the origin of the regulation of A, B, and C genes by LFY, a central part of the network leading to the emergence and development of flowers (Theissen and Melzer, 2007; Moyroud et al., 2010).

METHODS

Plant Materials

Wild-type plants were of the Columbia-0 accession. *35S:LFY* has been described before (Nilsson et al., 1998). Seedlings were grown under long-day photoperiods at 23°C on Murashige and Skoog plates.

Systematic Evolution of Ligands by Exponential Enrichment

Selection Cycles

In vitro selection of aptamers was performed with fluorescent 81-mers and a recombinant version of the DNA binding domain of *Arabidopsis thaliana* LFY protein (LFY DBD) produced and purified as previously described (Hamès et al., 2008)

Initially, a random sequence library was synthesized by PCR amplification (98°C for 1 min and 30 s followed by 20 cycles of 98°C for 10 s, 55°C for 25 s, and 72°C for 15 s) with Phusion DNA polymerase (Ozyme) using 81-mers [5′-TGGAGAAGAGGAGAGATCTAGC$(N)_{30}$CTCTAGATCTTGT-TCTTCTTCGATTCCGG-3′] as template with a fluorescent forward primer (SElex-F, TAMRA 5′-TGGAGAAGAGGAGAGATCTAG-3′) and a nonlabeled reverse primer (SElex-R, 5′-CCGGAATCGAAGAAGAACAA-3′) (Sigma-Aldrich). The size of the PCR products was verified on 3% agarose gels stained with SYBR Safe (Invitrogen), and double-stranded DNA (dsDNA) concentration was measured using SYBR green (Invitrogen) and a microplate reader (Safire^2; TECAN) according to the manufacturer's instructions.

For each selection cycle, 200 nM LFY-C was mixed to 10 nM fluorescent dsDNA (81-mers) in 225 μL Selex buffer (20 mM Tris, pH 8, 250 mM NaCl, 2 mM $MgCl_2$, 5 mM TCEP, 10 μg/mL dIdC, and 1% glycerol). After a 2-min incubation on ice, 25 μL Ni Sepharose 6 fast flow (GE Healthcare),

previously equilibrated in Selex buffer without TCEP, was added to the reaction mix to immobilize the DNA/protein complexes via the His tag of the protein. After 30 min incubation at 4°C on a rotating wheel, the reaction mix was loaded on an Ultrafree-MC centrifugal filter unit (Millipore) and centrifuged for 1 min at 500*g* at 4°C to eliminate the unbound DNA. Four washes were subsequently made by adding 300 μL of Selex buffer without dIdC on top of the filter unit followed by 1 min centrifugation at 500*g* at 4°C. Finally, the Ni Sepharose was resuspended in 100 μL water and transferred into a clean tube. Selected 81-mers were amplified by PCR as described above, using 2 μL of the Ni-Sepharose solution as template. PCR products were quantified as described before, and the selection cycle was repeated seven times, using each time the newly synthesized fluorescent DNA as a library.

The whole selection process has been performed twice independently.

Enrichment Evaluation

An electrophoretic mobility shift assay (Hamès et al., 2008) was used to estimate the enrichment for 81-mers with a high affinity for LFY DBD through the successive selection cycles: 10 nM 81-mers library of each cycle was incubated with 200 nM LFY DBD in 20 μL binding buffer. Electrophoresis and gel analysis was performed as described for QuMFRA assays, and libraries that gave a visible shift were selected for sequencing (cycles 3 to 7) using the 454 technology (Cogenics). More than 2500 sequences were obtained.

These sequences yielded 494 unique sequences, which were aligned with the MEME software version 4.3.0 (Bailey and Elkan, 1994) (http://meme.sdsc.edu/meme4_3_0/cgi-bin/meme.cgi) using the default parameters with either no constraints or with the symmetry imposed. This alignment was subsequently analyzed with the enoLOGOS software to identify dependence between nucleotides (Workman et al., 2005). For PSSM generation, frequencies of individual nucleotides and/or triplets were derived from the alignments and used to calculate, at each position *i* of the motif, the weight (*W*) associated to each nucleotide (or triplet) *n* according to: $W_{n,i} = \ln(f_{n,i}/f_{\max,i})$, where $f_{n,i}$ is the frequency of nucleotide *n* at position *i*, and $f_{\max,i}$ is the maximal frequency observed at position *i*. When $f_{n,i} = 0$, a pseudocount value (Wasserman and Sandelin, 2004) of 0.001 was applied.

QuMFRA Assay

QuMFRA assays were performed as described by Liu and Stormo (2005). Complementary single-stranded oligonucleotides were annealed in annealing buffer (10 mM Tris, pH 7.5, 150 mM NaCl, and 1 mM EDTA). The resulting dsDNA with a protruding G was fluorescently labeled by end-filling: 4 pmol of dsDNA was incubated with 1 unit of Klenow fragment polymerase (Ozyme) and 8 pmol Cy5-dCTP (GE Healthcare) (dsDNA samples) or Cy3-dCTP (dsDNA reference) in 1× Klenow buffer during 2 h at 37°C, followed by 10 min enzyme inactivation at 65°C. Sequences used as references or as samples are listed in Supplemental Table 4 online.

Binding reactions were performed in 20 μL binding buffer (20 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1% glycerol, 0.25 mM EDTA, 2 mM $MgCl_2$, 28 ng/mL fish sperm DNA [Roche], and 1 mM DTT) using 10 nM Cy3-dsDNA, 10 nM to 30 nM Cy5-dsDNA, and 500 nM or 1 μM LFY DBD. After 10 min incubation on ice, the binding reactions were loaded onto native 6% polyacrylamide gels and 0.5× TBE (45 mM Tris, 45 mM boric acid, and 1 mM EDTA, pH 8) and electrophoresed at 90 V for 90 min at 4°C.

Gels were scanned on a Typhoon 9400 scanner (Molecular Dynamics), and signals were quantified using ImageQuant software (Molecular Dynamics). Relative dissociation constants were calculated according to Man and Stormo (2001): for each gel lane, the fluorescent intensities of the bound and unbound fractions at both emission wavelengths were quantified and the background signal was subtracted. The resultant fluorescence intensities (FI_{cor}) were used to calculate the relative dissociation constant (K_D^{Rel}) given by Equation (1):

$$K_D^{Rel} = \frac{[FI_{cor}(\text{Bound})/FI_{cor}(\text{Free})]_{\text{reference}}}{[FI_{cor}(\text{Bound})/FI_{cor}(\text{Free})]_{\text{sample}}} \quad (1)$$

The relative dissociation constant of each dsDNA was measured at least three times independently, and the average value was used as K_D^{Rel} for comparison to the scores.

Experimental scores from Figures 1C to 1E are defined as $\ln(K_D^{Rel}/K_D^{Rel}\text{max})$, with $K_D^{Rel}\text{max}$ corresponding to K_D^{Rel} of the dsDNA with the highest affinity for LFY DBD.

Cross-Linking, Chromatin Isolation, and ChIP-seq

The entire experiment from seed sowing through deep sequencing was performed twice to produce independent biological replicates. ChIP-seq (Yant et al., 2010) was performed with an antibody raised in rabbit (#4028) against the LFY C-terminal amino acids 223 to 424 (BioGenes). Briefly, 15-d-old *35S:LFY* and Columbia-0 (control) seedlings were harvested and fixed as described previously (Gomez-Mena et al., 2005). Frozen tissue was ground, filtered three times through Miracloth (Calibrochem), and washed as described previously with buffers M1, M2, and M3 (Gomez-Mena et al., 2005). Nuclear pellets were resuspended in sonic buffer as described (1 mM PEFA BLOC SC [Roche Diagnostics] was substituted for PMSF), split into technical duplicate samples, and sonicated with a Branson sonifier at continuous pulse (output level 3) for eight rounds of 2 × 6 s and allowed to cool on ice between rounds. Immunoprecipitation reactions were performed by incubating chromatin with 2.5 μL anti-LFY serum overnight at 4°C as described (Gomez-Mena et al., 2005). The immunoprotein-chromatin complexes were captured by incubating with protein A-agarose beads (Santa Cruz Biotechnology), followed by consecutive washes in immunoprecipitation buffer and then elution as described (Gomez-Mena et al., 2005). Immunoprotein-DNA was then incubated consecutively in RNase A/T1 mix (Fermentas) and Proteinase K (Roche Diagnostics) as described after which DNA was purified using Minelute columns (Qiagen) (Gomez-Mena et al., 2005). ChIP samples were tested for enrichment by quantitative PCR, and deep sequencing libraries were produced by standard Illumina protocols.

ChIP-seq Analysis

Standard Illumina base calling software was used to base call the 40- to 42-nucleotide sequence reads. We used SHORE (Ossowski et al., 2008) as a platform for further analysis. The obtained reads were quality filtered, and low-quality bases at the 3′ end were pruned as described (Ossowski et al., 2008). GenomeMapper (Schneeberger et al., 2009) was used for mapping to the TAIR9 genome, allowing for up to four mismatching nucleotides and no gaps.

To proceed, the mapped data were subjected to a heuristic for removal of duplicate sequence reads, which were assumed to be uninformative for the detection of enriched loci. A threshold was applied limiting the number of 5′ ends mapping to the same position on the same strand. To retain the power to discriminate between multiple strongly enriched regions, the threshold for any particular position was varied depending on the coverage in close vicinity, such that the variance of the number of reads per position would roughly equal its mean in a 30-bp sliding window.

We further applied a two-step procedure to identify regions significantly enriched in the positive sample when compared with the control. First, potentially enriched regions were identified based on the positive samples only. These sites were then directly compared with the corresponding control sample regions to assess statistical significance.

For estimation of the depth of coverage for each position in the genome, all positive sample reads mapping to unique positions were extended in 3′ direction to 130 bp, corresponding to half the experimentally observed approximate DNA fragment size, while discarding all other reads. To detect possible peak sites, a 2-kb wide sliding window was applied to the coverage graph in single base steps. In each step a P value was assigned to the coverage value at the central base using a one-sided Poisson test, with the distribution parameter set to the average coverage within the sliding window. Only positions with coverage >0 were included in the calculation of the average, assuming all other positions to be inaccessible to the experiment. Finally, any consecutive stretch of positions with P value <0.05 and length >130 bp was retained as a potentially enriched site. To reduce further the number of regions to be considered, each was checked for unwarranted high average coverage in the control sample. A potential peak in the positive sample was discarded if the coverage mean in the control sample in the corresponding region was larger than the median average control coverage plus a tolerance of three standard deviations in all peak regions.

For assignment of final P values to each candidate region, in each replicate a one-sided binomial test was applied to the number of reads mapping to the region in the positive sample, with the distribution parameter N set to the joint read count for the site for the positive and the corresponding control samples. To estimate the probability parameter for the test, from now on called r, we computed a scaling factor s for the control sample and the chromosome containing the considered region. The complete chromosome sequence was subdivided into 400-bp bins, and for each bin, the positive sample and the control sample read counts were recorded. Then, s was chosen such that the median ChIP sample read count for all bins equaled the median control sample read count multiplied by s. From this the binomial test parameter, r was calculated as $r = s/(s + 1)$.

Finally, FDRs were obtained through the Benjamini-Hochberg correction method. To establish a ranking of peak regions across replicates, the rank product over the per-replicate FDR ranks was used.

Biophysical Model for LFY-DNA Binding

We used POcc (Roider et al., 2007), defined as the expected number of bound TF molecules for a given TF matrix of length W and a DNA sequence of length L, as given by Equation (2), where $K_{A,s}$ is the relative equilibrium association constant for sites.

$$POcc = \sum_{s=1}^{L-W} p_s = \sum_{s=1}^{L-W} \frac{K_{A,s} \cdot [\mathrm{TF}]}{1 + K_{A,s} \cdot [\mathrm{TF}]} \tag{2}$$

$K_{A,s}$ is the inverse of the relative equilibrium dissociation constant ($1/K_{D,s}$) and was calculated thanks to the correlation curve in Figure 1, as given by Equation 3:

$$\mathrm{score}_s = -\ln(K_{D,s})\, a + b \rightarrow K_{D,s} = e^{\frac{(b - \mathrm{score}_s)}{a}} \tag{3}$$

We found that $a = 1.6349$ and $b = -3.9647$ for the ASY PSSM, $a = 1.8031$ and $b = 0.4133$ for the SYM PSSM, and $a = 2.5663$ and $b = 0.3598$ for the SYM-T PSSM, and we used [TF] equal to the K_D for the optimal site (score = 0), resulting in $p_{s\text{-opt}} = 0.5$ (Granek and Clarke, 2005; Roider et al., 2007).

In the analyses presented in Figures 3 and 5, we used a variant of POcc in which only binding sites with a score higher than a threshold $t = -23$ are considered (Roider et al., 2007).

POcc was calculated for all peaks in ChIP experiment (~20,000). The correlation between ChIP and POcc ranking while using different PSSM was measured with the Spearman's rank correlation coefficient. This is a nonparametric measure of statistical dependence between the two variables ChIP and POcc. First, the n raw values (ChIPi and Pocci) were converted to ranks (x_i and y_i). Second, the differences, $d_i = x_i - y_i$, between the ranks of each observation on the two variables were calculated. The Spearman's rho (i.e., the correlation coefficient) was then given by Equation 4:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{4}$$

Selection of Bound Peaks Set and Unbound Genomic Set

To perform ROC analysis, the bound DNA set was composed of all peaks with FDR < 0.1 in both ChIP experiments, resulting in 1564 peaks. The peaks were ranked using the rank product from both ChIP-seq replicates. The unbound set was generated by randomly selecting 1564 sequences from the *A. thaliana* genome that did not overlap with bound fragments and with the same size distribution as the bound set.

Data Processing

Various scripts in Python (www.python.org; v2.6.4) were written for automatic data processing, including PSSM score calculation, POcc determination, and ROC-AUC estimation.

Microarray Data Source

Microarray data was retrieved from Gene Expression Omnibus data sets (www.ncbi.nlm.nih.gov/geo): record GDS515 (William et al., 2004) and record GDS453 (Schmid et al., 2003). From GDS453, we used wild-type plants versus *lfy12* floral transition microarrays at 0, 3, 5, and 7 d. From GDS515, we used dexamethasone versus mock treatment and dexamethasone+cycloheximide versus cycloheximide treatment in 35S:LFY-GR plants to select for potential direct targets of LFY. We selected all genes with a fold change higher than 2 in one of the conditions without attempting to calculate a statistical significance of this fold change.

The significance of the overlap between deregulated genes in the GDS515 microarray and the bound genes from the LEAFY ChIP-seq experiment was computed using a hypergeometric distribution, given by Equation 5:

$$\rho - \mathrm{value} = 1 - \sum_{x=0}^{k} P(X = x) = 1 - \sum_{x=0}^{k} \frac{\binom{M}{x}\binom{T-M}{N-x}}{\binom{T}{N}} \tag{5}$$

where M is the number of bound genes, N the number of deregulated genes in the microarray, T the total number of genes in the microarray, and k the number of genes that are both bound and deregulated. All computations were done using R software, and scripts are available upon request.

Genomic Sequence Retrieval and Analysis

For all species (except *A. thaliana*, *Antirrhinum majus*, *Brachypodium distachyon*, and *Sorghum bicolor*), the coding regions of previously identified members of the *AG* subfamily (see Supplemental Table 2 online for accession numbers) were retrieved from GenBank (http://www.ncbi.nlm.nih.gov) and used as BLAST queries against their respective species genome assembly to identify the corresponding genomic sequences. Coding sequences of members of the *AG* subfamily in *Oryza sativa* or *Zea mays* were blasted against the genomes of *S. bicolor* or *B. distachyon* to find the orthologs in these species. Plant genomes assemblies of *A. thaliana*, *Arabidopsis lyrata*, *Populus trichocarpa*, *Carica papaya*, *Vitis*

vinifera, *Prunus persica*, *Cucumis sativus*, *B. distachyon*, *O. sativa*, *S. bicolor*, and *Z. mays* were browsed and queried at Phytozome v5.0 (http://www.phytozome.net). The *S. lycopersicum* genome assembly (v1.50) was browsed and queried at the Sol genomic network (http://solgenomics.net). The POcc values (t = −23) were then calculated on the longest intron of each gene, which corresponds to the first or the second intron depending on the gene. The accession numbers for the large intron of *AG* orthologs in Brassicaceae (Hong et al., 2003) can be found on Supplemental Table 5 online.

Intron sequences were aligned with DIALIGN software (Morgenstern, 2004), and a sliding-window analysis with a window size of 20 bp was used to estimate the mean divergence between sequences using the Jukes-Cantor model. The inverse of the mean divergence (mean conservation) is represented on Supplemental Figures 2 to 4 online.

Accession Numbers

All ChIP-seq data are freely available from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/; accession number GSE24568). Sequence data from this article can be found in the Arabidopsis Genome Initiative or GenBank/EMBL databases under the following accession numbers: AY935269 (*PLE*), AY935268 (*FARINELLI*), AT4G18960 (*AG*), AT3G58780 (*SHP1*), AT2G42830 (*SHP2*), and AT4G09960 (*STK*). All other accession numbers are listed in Table 1 and Supplemental Tables 2 and 5 online.

Author Contributions

F.P., E.M., M.S., and L.Y. designed the experiments, and E.M., L.Y., D.P., S.B., and E.T. performed the experiments. E.G.M. and F.O. performed the data analysis with contributions from E.M., O.B., and M.M. The article was written by F.P., M.S, and D.W. with contributions from E.M. and L.Y.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Example of LFY-Bound Regions Identified by ChIP-seq and Analyzed for LFY Binding Sites.

Supplemental Figure 2. LFY Binding Sites in the Large Intron of *AG* Homologs in Eudicots.

Supplemental Figure 3. LFY Binding Sites in the Large Intron of *AG* Homologs in Monocots.

Supplemental Figure 4. LFY Binding Sites in the Second Intron of *AG* Homologs in Brassicaceae.

Supplemental Table 1. Position-Specific Scoring Matrix.

Supplemental Table 2. References for Data on the Expression of Genes of the *AG* Subfamily in Various Plant Species.

Supplemental Table 3. Predictions of LFY Binding at the Genomic Level.

Supplemental Table 4. Sequences of Oligonucleotides Used in LFY-DNA Interaction Studies.

Supplemental Table 5. Accession Numbers Corresponding to the Sequences of the Large Intron of *AG* Homologs in Brassicaceae Species.

Supplemental Data Set 1A. List of the 1564 Regions Bound by LFY in ChIP-seq Experiments.

Supplemental Data Set 1B. List of Genes (Upstream and Downstream) Adjacent to the 1564 Bound Regions in ChIP-seq Experiments.

Supplemental Data Set 1C. Overlap between Genes Bound by LFY in ChIP-seq and Genes Regulated by LFY.

ACKNOWLEDGMENTS

We thank C. Scutt, P. Lemaire, R. Vincentelli, K. Nitta, and members of the Parcy and Schmid laboratories for discussion and A.K. Martin for help with bioinformatic analyses. This work was supported by funding from the Centre National de la Recherche Scientifique (ATIP+; F.P.), the Agence Nationale de la Recherche (ANR, Plant-TFcode; F.P.), the ANR and the Biotechnology and Biological Sciences Research Council (Flower Model; F.P.), and PhD fellowships from the University J. Fourier, Grenoble (E.M. and M.M.), FP7 Collaborative Project AENEAS (Contract KBBE-2009-226477; D.W.), ERA-NET Plant Genomics Project BLOOM-NET (SCHM 1560/7-1; M.S.), and the Max Planck Society (M.S. and D.W.).

Received January 24, 2011; revised March 22, 2011; accepted April 1, 2011; published April 22, 2011.

REFERENCES

Badis, G., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. Science **324:** 1720–1723.

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. **2:** 28–36.

Benos, P.V., Bulyk, M.L., and Stormo, G.D. (2002). Additivity in protein-DNA interactions: How good an approximation is it? Nucleic Acids Res. **30:** 4442–4451.

Bomblies, K., Wang, R.L., Ambrose, B.A., Schmidt, R.J., Meeley, R.B., and Doebley, J. (2003). Duplicate *FLORICAULA/LEAFY* homologs *zfl1* and *zfl2* control inflorescence architecture and flower patterning in maize. Development **130:** 2385–2395.

Busch, M.A., Bomblies, K., and Weigel, D. (1999). Activation of a floral homeotic gene in *Arabidopsis*. Science **285:** 585–587.

Causier, B., Bradley, D., Cook, H., and Davies, B. (2008). Conserved intragenic elements were critical for the evolution of the floral C-function. Plant J. **1:** 41–52.

Causier, B., Castillo, R., Zhou, J., Ingram, R., Xue, Y., Schwarz-Sommer, Z., and Davies, B. (2005). Evolution in action: Following function in duplicated floral homeotic genes. Curr. Biol. **15:** 1508–1512.

Colombo, M., Brambilla, V., Marcheselli, R., Caporali, E., Kater, M.M., and Colombo, L. (2010). A new role for the *SHATTERPROOF* genes during *Arabidopsis* gynoecium development. Dev. Biol. **337:** 294–302.

Davies, B., Motte, P., Keck, E., Saedler, H., Sommer, H., and Schwarz-Sommer, Z. (1999). PLENA and FARINELLI: Redundancy and regulatory interactions between two Antirrhinum MADS-box factors controlling flower development. EMBO J. **18:** 4023–4034.

Dowell, R.D. (2010). Transcription factor binding variation in the evolution of gene regulation. Trends Genet. **26:** 468–475.

Ferrario, S., Immink, R.G., and Angenent, G.C. (2004). Conservation and diversity in flower land. Curr. Opin. Plant Biol. **7:** 84–91.

Gan, Y., Kumimoto, R., Liu, C., Ratcliffe, O., Yu, H., and Broun, P. (2006). *GLABROUS INFLORESCENCE STEMS* modulates the regulation by gibberellins of epidermal differentiation and shoot maturation in *Arabidopsis*. Plant Cell **18:** 1383–1395.

Gomez-Mena, C., de Folter, S., Costa, M.M., Angenent, G.C., and

Sablowski, R. (2005). Transcriptional program controlled by the floral homeotic gene AGAMOUS during early organogenesis. Development **132:** 429–438.

Granek, J.A., and Clarke, N.D. (2005). Explicit equilibrium modeling of transcription-factor binding and gene regulation. Genome Biol. **6:** R87.

Hamès, C., Ptchelkine, D., Grimm, C., Thevenon, E., Moyroud, E., Gérard, F., Martiel, J.L., Benlloch, R., Parcy, F., and Müller, C.W. (2008). Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. EMBO J. **27:** 2628–2637.

Hanley, J.A., and McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **143:** 29–36.

Hong, R.L., Hamaguchi, L., Busch, M.A., and Weigel, D. (2003). Regulatory elements of the floral homeotic gene *AGAMOUS* identified by phylogenetic footprinting and shadowing. Plant Cell **15:** 1296–1309.

Kasowski, M., et al. (2010). Variation in transcription factor binding among humans. Science **328:** 232–235.

Kaufmann, K., Wellmer, F., Muiño, J.M., Ferrier, T., Wuest, S.E., Kumar, V., Serrano-Mislata, A., Madueño, F., Krajewski, P., Meyerowitz, E.M., Angenent, G.C., and Riechmann, J.L. (2010). Orchestration of floral initiation by *APETALA1*. Science **328:** 85–89.

Kramer, E.M., Jaramillo, M.A., and Di Stilio, V.S. (2004). Patterns of gene duplication and functional evolution during the diversification of the *AGAMOUS* subfamily of MADS box genes in angiosperms. Genetics **166:** 1011–1023.

Lamb, R.S., Hill, T.A., Tan, Q.K., and Irish, V.F. (2002). Regulation of *APETALA3* floral homeotic gene expression by meristem identity genes. Development **129:** 2079–2086.

Liljegren, S.J., Ditta, G.S., Eshed, Y., Savidge, B., Bowman, J.L., and Yanofsky, M.F. (2000). *SHATTERPROOF* MADS-box genes control seed dispersal in *Arabidopsis.* Nature **404:** 766–770.

Liljegren, S.J., Gustafson-Brown, C., Pinyopich, A., Ditta, G.S., and Yanofsky, M.F. (1999). Interactions among *APETALA1*, *LEAFY*, and *TERMINAL FLOWER1* specify meristem fate. Plant Cell **11:** 1007–1018.

Liu, C., Thong, Z., and Yu, H. (2009). Coming into bloom: The specification of floral meristems. Development **136:** 3379–3391.

Liu, J., and Stormo, G.D. (2005). Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions. Nucleic Acids Res. **33:** e141.

Lohmann, J.U., Hong, R.L., Hobe, M., Busch, M.A., Parcy, F., Simon, R., and Weigel, D. (2001). A molecular link between stem cell regulation and floral patterning in *Arabidopsis*. Cell **105:** 793–803.

Man, T.K., and Stormo, G.D. (2001). Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. Nucleic Acids Res. **29:** 2471–2478.

Mikkelsen, T.S., Xu, Z., Zhang, X., Wang, L., Gimble, J.M., Lander, E.S., and Rosen, E.D. (2010). Comparative epigenomic analysis of murine and human adipogenesis. Cell **143:** 156–169.

Morgenstern, B. (2004). DIALIGN: Multiple DNA and protein sequence alignment at BiBiServ. Nucleic Acids Res. **32**(Web Server issue): W33–W36.

Moyroud, E., Kusters, E., Monniaux, M., Koes, R., and Parcy, F. (2010). LEAFY blossoms. Trends Plant Sci. **15:** 346–352.

Moyroud, E., Tichtinsky, G., and Parcy, F. (2009). The LEAFY floral regulators in Angiosperms: Conserved proteins with diverse roles. J. Plant Biol. **52:** 177–185.

Nilsson, O., Lee, I., Blázquez, M.A., and Weigel, D. (1998). Flowering-time genes modulate the response to LEAFY activity. Genetics **150:** 403–410.

Odom, D.T., Dowell, R.D., Jacobsen, E.S., Gordon, W., Danford, T.W., MacIsaac, K.D., Rolfe, P.A., Conboy, C.M., Gifford, D.K., and Fraenkel, E. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat. Genet. **39:** 730–732.

Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N., and Weigel, D. (2008). Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res. **18:** 2024–2033.

Parcy, F., Bomblies, K., and Weigel, D. (2002). Interaction of *LEAFY*, *AGAMOUS* and *TERMINAL FLOWER1* in maintaining floral meristem identity in *Arabidopsis*. Development **129:** 2519–2527.

Parcy, F., Nilsson, O., Busch, M.A., Lee, I., and Weigel, D. (1998). A genetic framework for floral patterning. Nature **395:** 561–566.

Ratcliffe, O.J., Bradley, D.J., and Coen, E.S. (1999). Separation of shoot and floral identity in *Arabidopsis*. Development **126:** 1109–1120.

Roider, H.G., Kanhere, A., Manke, T., and Vingron, M. (2007). Predicting transcription factor affinities to DNA from a biophysical model. Bioinformatics **23:** 134–141.

Schmid, M., Uhlenhaut, N.H., Godard, F., Demar, M., Bressan, R., Weigel, D., and Lohmann, J.U. (2003). Dissection of floral induction pathways using global expression analysis. Development **130:** 6001–6012.

Schmidt, D., et al. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science **328:** 1036–1040.

Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O., and Weigel, D. (2009). Simultaneous alignment of short reads against multiple genomes. Genome Biol. **10:** R98.

Sieburth, L.E., and Meyerowitz, E.M. (1997). Molecular dissection of the *AGAMOUS* control region shows that cis elements for spatial regulation are located intragenically. Plant Cell **9:** 355–365.

Theissen, G., and Melzer, R. (2007). Molecular mechanisms underlying origin and diversification of the angiosperm flower. Ann. Bot. (Lond.) **100:** 603–619.

Thompson, B.E., and Hake, S. (2009). Translational biology: From *Arabidopsis* flowers to grass inflorescence architecture. Plant Physiol. **149:** 38–45.

Wagner, D., Sablowski, R.W., and Meyerowitz, E.M. (1999). Transcriptional activation of *APETALA1* by LEAFY. Science **285:** 582–584.

Ward, L.D., and Bussemaker, H.J. (2008). Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. Bioinformatics **24:** i165–i171.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. **5:** 276–287.

Weirauch, M.T., and Hughes, T.R. (2010). Conserved expression without conserved regulatory sequence: The more things change, the more they stay the same. Trends Genet. **26:** 66–74.

Whitington, T., Perkins, A.C., and Bailey, T.L. (2009). High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. Nucleic Acids Res. **37:** 14–25.

William, D.A., Su, Y., Smith, M.R., Lu, M., Baldwin, D.A., and Wagner, D. (2004). Genomic identification of direct target genes of LEAFY. Proc. Natl. Acad. Sci. USA **101:** 1775–1780.

Wilson, M.D., and Odom, D.T. (2009). Evolution of transcriptional control in mammals. Curr. Opin. Genet. Dev. **19:** 579–585.

Won, K.J., Ren, B., and Wang, W. (2010). Genome-wide prediction of transcription factor binding sites using an integrated model. Genome Biol. **11:** R7.

Workman, C.T., Yin, Y., Corcoran, D.L., Ideker, T., Stormo, G.D., and Benos, P.V. (2005). enoLOGOS: A versatile web tool for energy

normalized sequence logos. Nucleic Acids Res. **33**(Web Server issue): W389–W392.

Yant, L., Mathieu, J., Dinh, T.T., Ott, F., Lanz, C., Wollmann, H., Chen, X., and Schmid, M. (2010). Orchestration of the floral transition and floral development in *Arabidopsis* by the bifunctional transcription factor APETALA2. Plant Cell **22:** 2156–2170.

Yeo, Z.X., Yeo, H.C., Yeo, J.K., Yeo, A.L., Li, Y., and Clarke, N.D. (2009). Inferring transcription factor targets from gene expression changes and predicted promoter occupancy. J. Comput. Biol. **16:** 357–368.

Zahn, L.M., Leebens-Mack, J.H., Arrington, J.M., Hu, Y., Landherr, L.L., dePamphilis, C.W., Becker, A., Theissen, G., and Ma, H. (2006). Conservation and divergence in the *AGAMOUS* subfamily of MADS-box genes: Evidence of independent sub- and neofunctionalization events. Evol. Dev. **8:** 30–45.

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y.V., Pellegrini, M., Goodrich, J., and Jacobsen, S.E. (2007). Whole-genome analysis of histone H3 lysine 27 trimethylation in *Arabidopsis*. PLoS Biol. **5:** e129.

Zhao, Y., Granas, D., and Stormo, G.D. (2009). Inferring binding energies from selected binding sites. PLoS Comput. Biol. **5:** e1000590.

Supplemental data. Moyroud et al. Plant Cell. (2011). 10.1105/tpc.111.083329

Supplemental Table 3: Predictions of LFY binding at the genomic level.

The *Arabidopsis* genome was scanned using the SYM-T PSSM to compute the score for all 19 bp sites as well as the POcc for all 150 bp regions. Predicted bound regions were generated by fusing overlapping 150 bp regions displaying a POcc higher than the threshold.

Part A displays the number of sites (column 2) with a score better than the threshold from column 1 present in the *Arabidopsis* genome and the subset of these sites (number in column 3 and proportion in column 4) present in the 1564 regions bound in the ChIP-seq experiments.

Part B displays the number of regions with a POcc better than a threshold (column 1) and the subset of these regions that overlaps with the 1564 bound regions.

A: genomic analysis of predicted binding sites			
Score threshold (Th)	**Number of 19-bp sites with a score > Th (Ng)**	**Number of sites from column 2 and present in the 1564 bound regions (Nb)**	**Nb/Ng**
-5	4	1	25.00%
-7	33	9	27.27%
-8	76	20	26.32%
-9	164	40	24.39%
-10	396	61	15.40%
-11	828	122	14.73%
-12	1781	199	11.17%
-13	3773	304	8.06%
-14	7635	463	6.06%
-15	14885	662	4.45%
B: genomic analysis of predicted bound regions			
POcc threshold (PTh)	**Number of regions with a POcc > PTh (Ng)**	**Number of regions from column 2 overlapping with the 1564 bound regions (Nb)**	**Nb/Ng**
0.04	90	25	27.78%
0.02	471	84	17.83%
0.01	2076	232	11.18%
0.005	9633	549	5.70%
0.003	28030	891	3.18%
0.002	59284	1200	2.02%

Developmental Cell
Article

LEAFY Target Genes Reveal Floral Regulatory Logic, *cis* Motifs, and a Link to Biotic Stimulus Response

Cara M. Winter,[1] Ryan S. Austin,[2] Servane Blanvillain-Baufumé,[3] Maxwell A. Reback,[1] Marie Monniaux,[4] Miin-Feng Wu,[1] Yi Sang,[1] Ayako Yamaguchi,[1] Nobutoshi Yamaguchi,[1] Jane E. Parker,[3] Francois Parcy,[4] Shane T. Jensen,[5] Hongzhe Li,[6] and Doris Wagner[1,*]

[1]Department of Biology, University of Pennsylvania School of Arts and Sciences, Philadelphia, PA 19104, USA
[2]Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, M5S 3B2 ON, Canada
[3]Department of Plant-Microbe Interactions, Max-Planck Institute for Plant Breeding Research, 50829 Cologne, Germany
[4]Laboratoire Physiologie Cellulaire et Végétale CNRS, iRTSV/CEA, INRA, UJF Grenoble I, Grenoble, France
[5]Department of Statistics, University of Pennsylvania Wharton School, Philadelphia, PA 19104, USA
[6]Department of Biostatistics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA
*Correspondence: wagnerdo@sas.upenn.edu
DOI 10.1016/j.devcel.2011.03.019

SUMMARY

The transition from vegetative growth to flower formation is critical for the survival of flowering plants. The plant-specific transcription factor LEAFY (LFY) has central, evolutionarily conserved roles in this process, both in the formation of the first flower and later in floral patterning. We performed genome-wide binding and expression studies to elucidate the molecular mechanisms by which LFY executes these roles. Our study reveals that LFY directs an elaborate regulatory network in control of floral homeotic gene expression. LFY also controls the expression of genes that regulate the response to external stimuli in *Arabidopsis*. Thus, our findings support a key role for LFY in the coordination of reproductive stage development and disease response programs in plants that may ensure optimal allocation of plant resources for reproductive fitness. Finally, motif analyses reveal a possible mechanism for stage-specific LFY recruitment and suggest a role for LFY in overcoming polycomb repression.

INTRODUCTION

Angiosperms or flowering plants are the most successful clade of plants representing nearly 90% of extant land plants. To reach the next generation, flowering plant meristems must cease formation of leaves or branches and initiate formation of reproductive structures, the flowers (Poethig, 2003; Steeves and Sussex, 1989). This requires large-scale alterations in the transcriptional program during the meristem identity transition (Moyroud et al., 2010). This transition also triggers the switch from biomass and resource accumulation in the leaves to allocation of these resources to seed formation. Despite their importance for agriculture and plant reproductive success, the underlying regulatory mechanisms coordinating these events remain to be fully elucidated.

Optimal timing of the meristem identity transition is of particular import in monocarpic (annual) plants such as *Arabidopsis thaliana*, which only flower once in their life. Hence this developmental switch is tightly controlled by both environmental signals such as daylength, temperature, and light (quantity and quality), and by endogenous cues including the age of the plant (Kim et al., 2009; Kobayashi and Weigel, 2007; Komeda, 2004; Turck et al., 2008). These pathways converge to upregulate the expression of meristem identity genes including the plant-specific transcription factor *LEAFY* (*LFY*) (Liu et al., 2009a; Parcy, 2005). LFY is necessary and sufficient for the correct induction of floral fate, and is considered a master regulator of the meristem identity transition (Blazquez et al., 2006; Moyroud et al., 2010; Weigel et al., 1992; Weigel and Nilsson, 1995). Subsequently, LFY directs floral organ patterning by activating floral homeotic gene expression (Krizek and Fletcher, 2005; Weigel and Meyerowitz, 1993).

To gain insight into the regulatory mechanisms coordinating reproductive development, we used chromatin immunoprecipitation coupled with tiling array hybridization (ChIP-chip) and transcriptional profiling to uncover the range of activities and direct transcriptional changes effected by LFY during the meristem identity transition and during flower development.

RESULTS

Genomic Regions Bound by LFY at Two Developmental Stages

First, we identified genes bound by LFY during the meristem identity transition in seedlings. Because of the low *LFY* levels present at this early stage (Blazquez et al., 1997) (see Figures S1A and S1B available online), we employed an inducible form of LFY, 35S:LFY-GR, which fully rescues the *lfy* null mutant phenotype (Wagner et al., 1999). We have previously shown that activation of 35S:LFY-GR in 9-day-old wild-type seedlings allows identification of direct LFY target genes with a role in the meristem identity transition (Saddic et al., 2006; Wagner et al., 1999; Pastore et al., submitted; William et al., 2004). After treating 9-day-old 35S:LFY-GR seedlings for 4 hr with

Developmental Cell
LEAFY Regulatory Targets

dexamethasone, we immunoprecipitated LFY-DNA complexes using anti-LFY antiserum (Wagner et al., 1999; William et al., 2004) and hybridized the associated DNA fragments to *Arabidopsis* whole-genome tiling arrays (Figure S1B). Using a moving average algorithm (Ji et al., 2008), we identified 1588 significant LFY binding peaks at a false discovery rate of <0.05. Independent validation of enrichment indicates that the FDR is likely lower (Figure S1C). The low signal in control immunoprecipitations, the narrow LFY binding peak width, and the high average ChIP enrichment provide additional evidence for the quality of the ChIP-chip data (Figure S1D). The 1588 binding peaks were associated with 1296 unique genes (see Supplemental Experimental Procedures for details). Six of the seven known direct LFY meristem identity targets were identified by ChIP-chip, including *APETALA1* (*AP1*) and *LATE MERISTEM IDENTITY* 1 (*LMI1*) (Figure 1A; Figure S1E) (Busch et al., 1999; Parcy et al., 1998; Saddic et al., 2006; Wagner et al., 1999; William et al., 2004).

In a second experiment, we identified genes bound by endogenous LFY during floral patterning in 19-day-old wild-type inflorescences bearing young flower primordia using anti-LFY antiserum for ChIP. This analysis uncovered a total of 867 significant LFY binding peaks (FDR < 0.05) and 748 associated unique genes. The quality of this ChIP-chip data set was equivalent to that obtained at the seedling stage (Figure S1). Both of the known floral homeotic LFY target genes *APETALA3* (*AP3)* and *AGAMOUS* (*AG)* (Busch et al., 1999; Lamb et al., 2002) were identified in the inflorescence ChIP-chip (Figure 1A). LFY bound to a promoter proximal region of *AP3* known to be important for proper expression in developing flower primordia and to the previously defined LFY-responsive enhancer in the second intron of *AG* (Busch et al., 1999; Hill et al., 1998).

Comparison of LFY target genes identified at both stages revealed a significant overlap ($p < 10^{-296}$), providing independent validation of a subset of the LFY targets (Figure 1B). This overlap is expected because LFY continues to induce floral fate in incipient primordia in inflorescences (Blazquez et al., 2006). Consistent with a possible role for LFY binding events in transcriptional regulation, binding peaks clustered near transcription start sites (Figure 1C; Table S1).

To determine how frequently LFY binding leads to rapid changes in gene expression, we used the same LFY-GR activation procedure as for ChIP followed by transcriptome analysis. Forty-one percent of the genes bound by LFY at the seedling stage showed rapid changes in gene expression after LFY-GR activation (FDR < 0.05; Figure 1D) with 59% of these gene expression changes being greater than 1.5-fold (Figure S1F). Accordingly, LFY binding increases the probability that a given gene will exhibit altered expression in response to LFY-GR activation ($p < 10^{-16}$, logistic regression; Figure 1E). Some of the remaining LFY targets may require longer periods of LFY induction to show significant expression changes, or may be regulated by LFY only after accumulation of a cofactor not present in our experimental conditions. We observed nearly equivalent roles for LFY in up- and downregulation of gene expression (Figure 1D), in agreement with previous reports that LFY can act as a transcriptional activator and repressor (Parcy et al., 2002; Wagner et al., 1999; William et al., 2004).

Selection of High-Confidence LFY Target Genes

Next, we identified a high-confidence list of likely physiologically relevant direct LFY target genes from the seedling and inflorescence ChIP-chip data sets using public transcriptome data (Schmid et al., 2003, 2005; Wellmer et al., 2006). Specifically, we selected LFY-bound genes that were significantly differentially expressed in *lfy* mutants relative to wild-type plants (FDR <0.05 and |fold change| >1.5), and that were also strongly coexpressed with endogenous *LFY* (Pearson correlation $p < 0.05$) (see Experimental Procedures for details; Figure S2 and Table S2). Relative to all *Arabidopsis* genes, LFY-bound genes were highly enriched for genes that met these criteria (Fisher's exact $p < 10^{-15}$; Table S2). About one-quarter of the seedling LFY target genes and of the inflorescence LFY target genes were LFY-dependent and *LFY*-coexpressed (Figure 1F). We used only these high-confidence seedling and inflorescence LFY target genes for further analyses.

LFY Controls Floral Homeotic Gene Expression via an Intricate Regulatory Network

To infer the predicted functions of the high-confidence LFY target genes, we performed Gene Ontology (GO) term analysis (see methods). This revealed strikingly different GO term enrichments ($p < 0.00005$) at the two developmental stages analyzed (Figure 1G). As expected, transcriptional regulators were significantly enriched among the target genes identified at both stages.

The most highly enriched GO terms for inflorescence LFY targets were "organ development" ($p < 10^{-14}$) and "flower development" ($p < 10^{-11}$), and all GO terms preferentially enriched at this stage were linked to cell fate specification, morphogenesis, and differentiation (Figure 1G). Accordingly, the list of high-confidence LFY targets in inflorescences included well-known developmental regulators (Table 1). For example, our studies identified the floral homeotic gene *PISTILLATA* (*PI*) as a high-confidence direct LFY target in inflorescences (Table 1 and Figure 2). Two additional genes, *SEPALLATA3* (*SEP3)*, which encodes a LFY cofactor (Liu et al., 2009b), as well as *EMBRYONIC FLOWER1* (*EMF1)*, which encodes a polycomb regulator (Calonje et al., 2008), were LFY-bound, -dependent, and -coexpressed at this stage (Figure 2A and Table 1). LFY bound to regions in the *PI* promoter previously shown to be important for proper expression of this gene (Honma and Goto, 2000). We detected strong LFY binding peaks in the promoter and in the first intron of *SEP3* (Figure 2A). A previous study showed that the *SEP3* intron is important for correct expression (de Folter et al., 2007). Finally, LFY was recruited to the 5′ UTR of *EMF1*.

To test whether LFY can indeed regulate expression of these genes, we employed a synchronous flower induction system (*ap1 cal* 35S:LFY-GR; Figure S3) (Wellmer et al., 2006). We observed rapid changes in expression of *PI*, *SEP3* and *EMF1* shortly after LFY-GR activation in *ap1 cal* inflorescences (Figure 2B). While *PI* and *SEP3* were upregulated, *EMF1* was repressed by LFY (Figure 2B). *EMF1* is a polycomb regulator thought to directly repress expression of floral homeotic genes outside of the proper developmental context (Calonje et al., 2008; Chen et al., 1997); hence, downregulation of *EMF1* expression may be a prerequisite for *AP3*, *PI,* and *AG* upregulation and flower patterning.

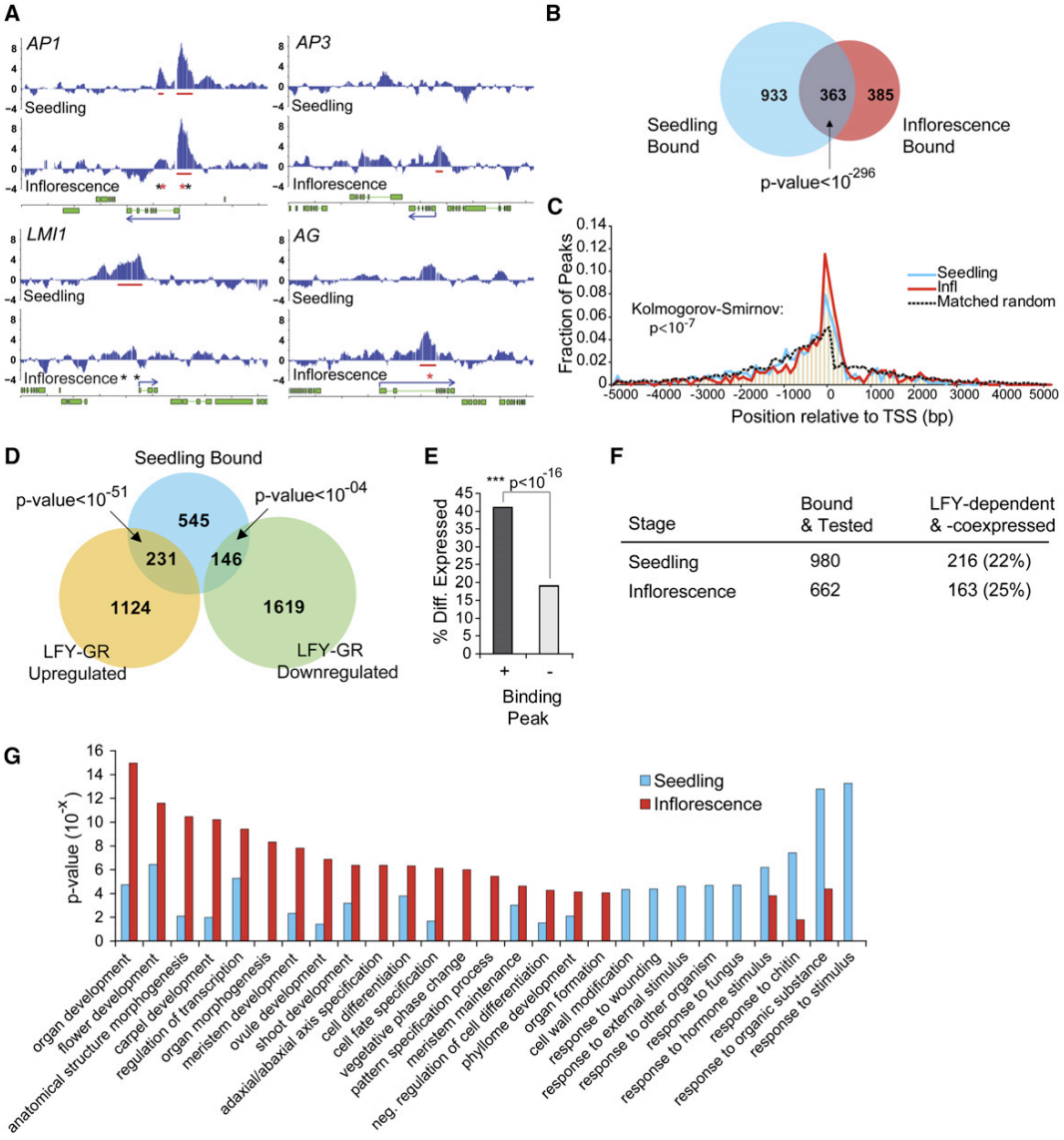


Figure 1. Genome-Wide LFY Binding to Regulatory Regions at Two Stages in Development

(A) Significant LFY binding at known direct LFY targets (Busch et al., 1999; Lamb et al., 2002; Wagner et al., 1999; William et al., 2004). Tracks: moving average t-statistic (20 kb window) for seedling (top) and inflorescence (bottom) ChIP-chip data. Horizontal red bars: significantly bound regions (FDR < 0.05). Asterisks: LFY consensus binding motifs in significantly bound regions ($p < 0.001$; red: primary; black: secondary; see text for details).

(B) Significant overlap (Fisher's exact test) between seedling and inflorescence LFY-bound target genes.

(C) LFY binding peaks map close to transcription start sites (TSSs). The pattern of LFY binding is significantly different from that of matched randomly generated peaks (70% intergenic, 30% genic; see Supplemental Experimental Procedures for details).

(D) Overlap of LFY-bound genes and genes differentially expressed (FDR < 0.05) in seedlings 4 hr after LFY-GR activation. Only 922 of 1296 LFY-bound genes were tested on the expression array (probe set is printed on array and passed our nonspecific filtering criteria).

(E) Presence of a seedling LFY binding peak significantly increased the probability that a gene was differentially expressed (logistic regression; $p < 10^{-16}$).

(F) Identification of high-confidence LFY-dependent and coexpressed LFY target genes (see text and Experimental Procedures for details). Nine hundred eighty seedling and 662 inflorescence targets were tested on the arrays used for the analysis.

(G) Gene ontology (GO) term enrichment ($p < 0.00005$ in at least one stage) for the high confidence LFY target genes. GO terms were grouped based on the stage of highest preferential enrichment and sorted based on p value.

See also Figures S1 and S2 and Tables S1 and S2.

The precise timing of the induction of the floral homeotic genes *AP3*, *PI*, and *AG* is central for proper flower morphogenesis; early induction leads to premature differentiation of the floral meristem, while late induction leads to the development of extra floral organs (Liu et al., 2009b). It was recently shown that this timing is critically linked to *SEP3* accumulation in the developing flower

Table 1. Examples of High-Confidence LFY Target Genes Identified at the Inflorescence Stage

	AGI ID	Gene Name	Role	Stage[a]	Citation[b]
Flowering time	AT1G53160	SPL4 (SQUA. PROMOTER BINDING PROTEIN-LIKE 4)	TXN	I	Amasino (2010); Michaels (2009)
	AT3G15270	SPL5 (SQUA. PROMOTER BINDING PROTEIN-LIKE 5)	TXN	I	
	AT1G72830	NF-YA8 /HAP2C	TXN	I S	
	AT5G47640	NF-YB2 /HAP3B	TXN	I	
	AT5G60910	FUL (FRUITFULL)	TXN	I	
Polarity	AT5G16560	KAN (KANADI)	TXN	I S	Bowman and Floyd (2008)
	AT1G32240	KAN2 (KANADI 2)	TXN	I	
	AT2G45190	FIL (FILAMENTOUS FLOWER)	TXN	I	
	AT2G34710	PHB (PHABULOSA)	TXN	I	
	AT2G37630	AS1 (ASYMMETRIC LEAVES 1)	TXN	I	
	AT5G60450	ARF4 (AUXIN RESPONSE FACTOR 4)	TXN	I	
	AT2G33860	ETT (ETTIN)	TXN	I	
Floral homeotic	AT3G54340	AP3 (APETALA 3)	TXN	I	Krizek and Fletcher (2005)
	AT5G20240	PI (PISTILLATA)	TXN	I	
	AT4G18960	AG (AGAMOUS)	TXN	I	
	AT1G24260	SEP3 (SEPALLATA 3)	TXN	I	
	AT2G03710	SEP4 (SEPALLATA 4)	TXN	I	
	AT5G11530	EMF1 (EMBRYONIC FLOWER 1)	CHR	I S	
Flower development	AT4G37750	ANT (AINTEGUMENTA)	TXN	I	Irish (2010); Jack (2004); van Zanten et al. (2009)
	AT5G10510	AIL6 (AINTEGUMENTA-LIKE 6)	TXN	I	
	AT5G35770	SAP (STERILE APETALA)	CHR	I	
	AT5G28640	AN3 (ANGUSTIFOLIA 3)	CHR	I	
	AT3G13960	GRF5 (GROWTH REGULATING FACTOR 5)	TXN	I	
	AT5G53950	CUC2 (CUP-SHAPED COTYLEDON 2)	TXN	I	
	AT4G36260	STY2 (STYLISH 2)	TXN	I	
	AT5G11320	YUC4 (YUCCA 4)	BS	I	
	AT1G70510	KNAT2 (KNOTTED-LIKE FROM HOMEOBOX GENE 2)	TXN	I	
	AT2G26330	ER (ERECTA)	SIG	I	
	AT5G62230	ERL1 (ERECTA-LIKE 1)	SIG	I S	
	AT5G07180	ERL2 (ERECTA-LIKE 2)	SIG	I	

AGI ID, locus identifier; BS, biosynthesis; CHR, chromatin-based regulation of transcription; TXN, transcription; SIG, signal transduction.
[a] Stage at which the direct LFY target gene was identified. S (seedling), I (inflorescence).
[b] Citation for functional grouping of target genes.

primordium (Kaufmann et al., 2009; Liu et al., 2009b). We therefore next investigated *SEP3* expression in *lfy* mutants compared to the wild-type using in situ hybridization and reporter studies (de Folter et al., 2007). In *lfy* mutants, we observed strongly reduced *SEP3* expression in the center of early stage 3 flower primordia (Figure 2C), the stage and tissue in which SEP3 upregulates the floral homeotic genes in wild-type plants (Liu et al., 2009b). Our data suggest that LFY directly induces the expression of its cofactor *SEP3* at this critical stage in flower development (Figure 2D).

LFY Moderates Biotic Stress Responses

High-confidence LFY target genes at the seedling stage were significantly enriched ($p < 10^{-4}$) in GO terms linked to development (Figure 1G) and included known regulators of the switch to reproduction, as expected (Table 2). Surprisingly, the majority of the GO terms enriched at this stage were associated with plant responses to endogenous (hormone) or environmental (biotic stress) stimuli (Figure 1G). Accordingly, known hormone and biotic stimulus response pathway regulators were among the identified high-confidence LFY targets (Table 2). Modulation of hormone response gene expression by LFY is consistent with roles for these pathways in primordium initiation and flower development (Liu et al., 2009a), while identification of biotic stimulus response genes as direct LFY target genes (Table 2) suggests a role for LFY in additional survival programs. Involvement of a developmental regulator in defense responses is not unprecedented (Nurmberg et al., 2007).

Two of the defense response LFY targets we identified, the ABC transporter *PDR8/PEN3* and the MAMP (microbe-associated molecular pattern) recognition receptor *FLS2*, were bound and repressed by LFY (Figures 3A and 3B). Both PEN3 and FLS2 are components of a basal plant immune response pathway leading to callose deposition at the cell wall and

Developmental Cell
LEAFY Regulatory Targets

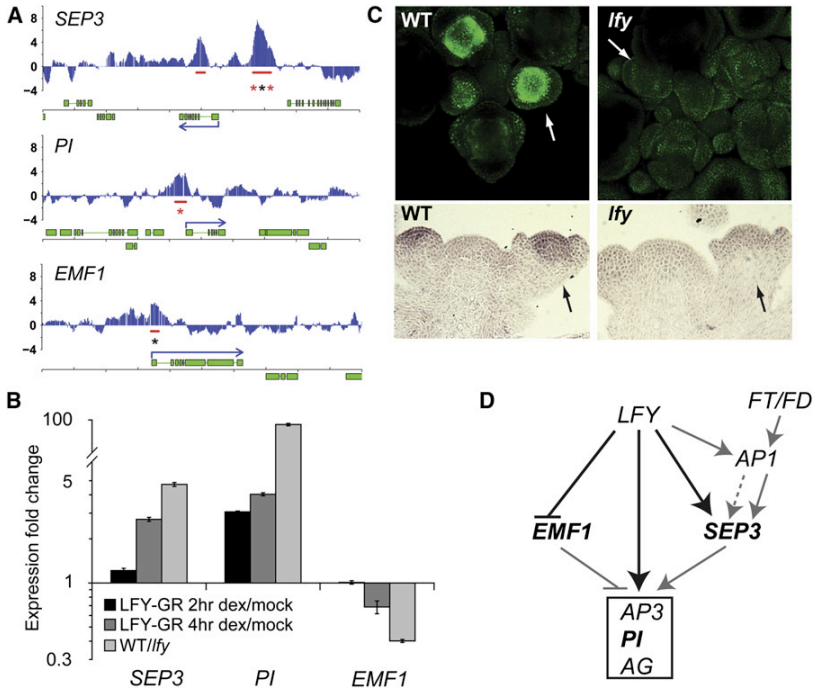


Figure 2. LFY Directly Regulates the Expression of Well-Known Developmental Regulators

(A) Significant LFY binding to regulatory regions of *SEP3*, *PI*, and *EMF1* in inflorescences (see Figure 1A for description of labels).

(B) Expression changes of these direct LFY targets after dexamethasone (dex) induction of LFY-GR in *ap1 cal* inflorescences relative to mock treated samples, and in wild-type (WT) compared with *lfy* null mutants in 13-day-old seedlings (*EMF1*) or in 15-day-old seedlings (*PI, SEP3*). Data shown are mean ± standard error of the mean (SEM).

(C) Top: Confocal images of wild-type and *lfy* null mutant inflorescences expressing pSEP3:SEP3-GFP (de Folter et al., 2007), which monitors LFY binding to the first intron of *SEP3*. Bottom: in situ hybridization of wild-type and *lfy* null mutant inflorescences showing expression of *SEP3*. Arrows point to young stage three flower primordia.

(D) Regulatory network controlling *AP3*, *PI*, and *AG* induction. Regulatory interactions identified here (black arrows and lettering) are supported by four independent criteria: the target gene is directly bound by LFY, coexpressed with *LFY*, and differentially expressed after LFY-GR activation and in *lfy* mutants compared with the wild-type. Dotted arrow: indirect regulation.

See Figure S3 for a comparison of LFY, AP1, and SEP3 binding data.

restriction of pathogen spread in the host (Clay et al., 2009; Zipfel et al., 2004). To investigate a possible link between LFY and this pathway, we challenged plants with a flagellin-derived peptide (flg22), which is recognized by FLS2. This yielded robust callose deposition in wild-type and in mock-treated LFY-GR seedlings, but not after prior steroid activation of LFY-GR (Figure 3C; Figures S4A and S4B).

To test for a role of endogenous LFY in this pathway, we examined callose deposits in flg22-infiltrated wild-type and *lfy* null mutant cauline leaves. *LFY* is expressed in this tissue during the meristem identity transition (Blazquez et al., 1997) (Figure S1A). We observed a significant increase in the number of flg22-induced callose deposits in *lfy* mutant relative to wild-type cauline leaves (Figure 3D). Consistent with this finding, many genes associated with this defense-induced cell wall modification pathway (Clay et al., 2009) were more highly expressed in this tissue in the *lfy* mutant than in the wild-type after flg22 treatment (Figure 3E). To probe additional FLS2-mediated downstream responses, we monitored the expression of flg22-induced defense genes not linked to callose deposition (Denoux et al., 2008). Upon flg22 stimulation, these genes also were more highly induced in *lfy* mutants than in the wild-type. Moreover, a gene encoding a lipid transfer protein inhibitor, known to be downregulated upon flg22 treatment, was more strongly repressed in *lfy* mutants (Figure 3F). Prolonged flg22 exposure inhibits plant growth (Gomez-Gomez et al., 1999). When we treated *lfy* mutant and wild-type seedlings for eight days with flg22 immediately after the meristem identity transition (in 11-day-old seedlings; Blazquez et al., 1997), the *lfy* seedlings exhibited more dramatic growth defects than the wild-type (Figure 3G). No significant difference in growth was observed between wild-type and *lfy* mutant seedlings when the treatment was initiated in younger seedlings (5-day-old; data not shown). Finally, we examined growth of a virulent bacterial strain (*Pseudomonas syringae* pv. *tomato* (*Pst*) DC3000) on wild-type and *lfy* mutant cauline and adult (late arising) rosette leaves; *LFY* is known to be expressed in the primordia of these leaves (Figure S1) (Blazquez et al., 1997). We observed a modest but significant decrease (3.5-fold, $p < 0.01$) in bacterial growth in the *lfy* mutant compared with the wild-type (Figure 3H; Figures S4C and S4D). Also, *lfy* mutant cauline leaves developed noticeably fewer disease symptoms than those of the wild-type (Figure 3H). These visible differences were not observed in *lfy* mutant rosette leaves (data not shown), consistent with the higher level of *LFY* expression in the later arising cauline leaf primordia (Blazquez et al., 1997). Wild-type and *lfy* mutant rosette leaves from 4-week-old short-day grown plants did not display differential defense gene expression, callose deposition, or pathogen growth when challenged with *Pst* DC3000 (Figures S4E–S4H), consistent with these leaves having formed prior to *LFY* induction (Blazquez et al., 1997; Hempel et al., 1997). Taken together, our results point to a role for LFY in reducing defense responses triggered by the MAMP flg22 and by bacterial pathogen challenge that may in part be attributable to downregulation of *FLS2* and *PEN3* levels by LFY.

De Novo Identification of Potential LFY Binding and Cofactor Motifs

The currently known LFY consensus binding motif, CCANTG[G/T], is based on only two experimentally confirmed LFY target genes (Busch et al., 1999). To better define a consensus LFY binding motif, we queried a subset of the seedling-and-inflorescence-bound

Developmental Cell
LEAFY Regulatory Targets

Table 2. Examples of High-Confidence LFY Target Genes Identified at the Seedling Stage

	AGI ID	Gene Name	Role	Stage[a]	Rapid DE[b]	Citation[c]
Flowering time	AT4G35900	FD	TXN	S	x	Amasino (2010); Michaels (2009); Yant et al. (2009)
	AT2G17770	FDP (FD PARALOG)	TXN	S I		
	AT1G27360	SPL11 (SQUA. PROMOTER BINDING PROTEIN-LIKE 11)	TXN	S		
	AT5G67180	TOE3 (TARGET OF EAT1 3)	TXN	S	x	
	AT1G25560	TEM1 (TEMPRANILLO 1)	TXN	S I	x	
Meristem identity	AT5G61850	LFY (LEAFY)	TXN	S I	NA	Albani and Coupland (2010); William et al. (2004)
	AT1G69120	AP1 (APETALA1)	TXN	S I	x	
	AT1G16070	AtTLP8/LMI5 (LATE MERISTEM IDENTITY 5)	SIG	S I	x	
	AT3G61250	MYB17/LMI2 (LATE MERISTEM IDENTITY 2)	TXN	S I	x	
	AT5G03840	TFL1 (TERMINAL FLOWER 1)	SIG	S I[d]	x	
Hormone response	AT2G34650	PID (PINOID)	T	S I	x	Bowman and Floyd (2008); Guilfoyle and Hagen (2007); Liu et al. (2009a)
	AT4G31820	ENP (ENHANCER OF PINOID); NPY1	T	S	x	
	AT5G67440	NPY3	T	S		
	AT2G01420	PIN4 (PIN-FORMED 4)	T	S		
	AT3G23050	IAA7 (AUXIN RESISTANT 2)	TXN	S		
	AT1G04250	AXR3 (AUXIN RESISTANT 3); IAA17	TXN	S I	x	
	AT2G33860	ETT (ETTIN); ARF3	TXN	S I		
	AT5G56300	GAMT2 (GIBBERELLIC ACID METHYLTRANSFERASE 2)	BS	S		Mutasa-Gottgens and Hedden (2009)
	AT3G63010	GID1B (GA INSENSITIVE DWARF1B)	RE	S	x	
Biotic stimulus response	AT2G39660	BIK1 (BOTRYTIS-INDUCED KINASE 1)	SIG	S		Burow et al. (2010); Clay et al. (2009); Dodds and Rathjen (2010)
	AT5G46330	FLS2 (FLAGELLIN-SENSITIVE 2)	SIG	S	x	
	AT1G59870	PDR8/PEN3 (PLEIOTROPIC DRUG RESISTANCE 8)	T	S		
	AT5G61420	MYB28 (MYB DOMAIN PROTEIN 28)	TXN	S	x	
	AT1G32540	LOL1 (LSD ONE LIKE 1)	TXN	S	x	
	AT2G38470	WRKY33 (WRKY DNA-BINDING PROTEIN 33)	TXN	S	x	
	AT1G80840	WRKY40 (WRKY DNA-BINDING PROTEIN 40)	TXN	S I		
	AT4G31800	WRKY18 (WRKY DNA-BINDING PROTEIN 18)	TXN	S	x	
	AT5G49520	WRKY48 (WRKY DNA-BINDING PROTEIN 48)	TXN	S I	x	

AGI ID, locus identifier; BS, biosynthesis; RE, receptor; SIG, signal transduction; T, transport; TXN, transcription.
[a] Stage at which the direct LFY target gene was identified. S (seedling), I (inflorescence).
[b] Genes significantly differentially expressed (FDR < 0.05) after 4 hr steroid activation of LFY-GR (expression array, see methods).
[c] Citation for functional grouping of target genes.
[d] LFY bound site in the 3′ intergenic region.

regions with a novel sequential analysis pipeline, which utilizes predictions from five popular motif-finding algorithms (see methods for details). We identified a 19 bp palindromic presumptive LFY binding motif, henceforth termed the "primary" LFY motif, that was strongly enriched ($p < 10^{-145}$) in all LFY-bound regions (Figure 4A; Figures S5A and S5B). A single motif prediction algorithm (Bailey and Elkan, 1995) identified a similar motif (Figure S5C). This primary LFY motif contained critical nucleotides contacted by a LFY DNA binding domain homodimer based on protein/DNA cocrystals (Hames et al., 2008) (Figure 4A). The previously identified CCANTG[G/T] consensus, while contained within the primary binding motif, was itself only marginally enriched (Figure 4A).

Many of the observed significant LFY binding events were specific to the seedling or the inflorescence data set (Figure 1B). For example, the known meristem identity regulator *LMI1* was bound by LFY only at the seedling stage, while the floral homeotic genes *AP3* and *AG* were bound only at the inflorescence stage (Figure 1A). To test for LFY binding motif variants that may contribute to this stage-specific LFY recruitment, we repeated our de novo motif analyses for a subset of regions bound only in inflorescences or only in seedlings. In inflorescences, we identified a motif similar to the primary LFY consensus motif (Figure S5A). Importantly, our analysis of seedling-only bound regions revealed a potential secondary LFY consensus motif, which was highly enriched in LFY-bound regions identified at the seedling stage ($p < 10^{-45}$) (Figures 4B and 4C; Figures S5A and S5B). This motif mainly differs from the primary LFY consensus at the +2 position relative to the motif core with a thymine preferred to guanine. A similar secondary

Developmental Cell
LEAFY Regulatory Targets

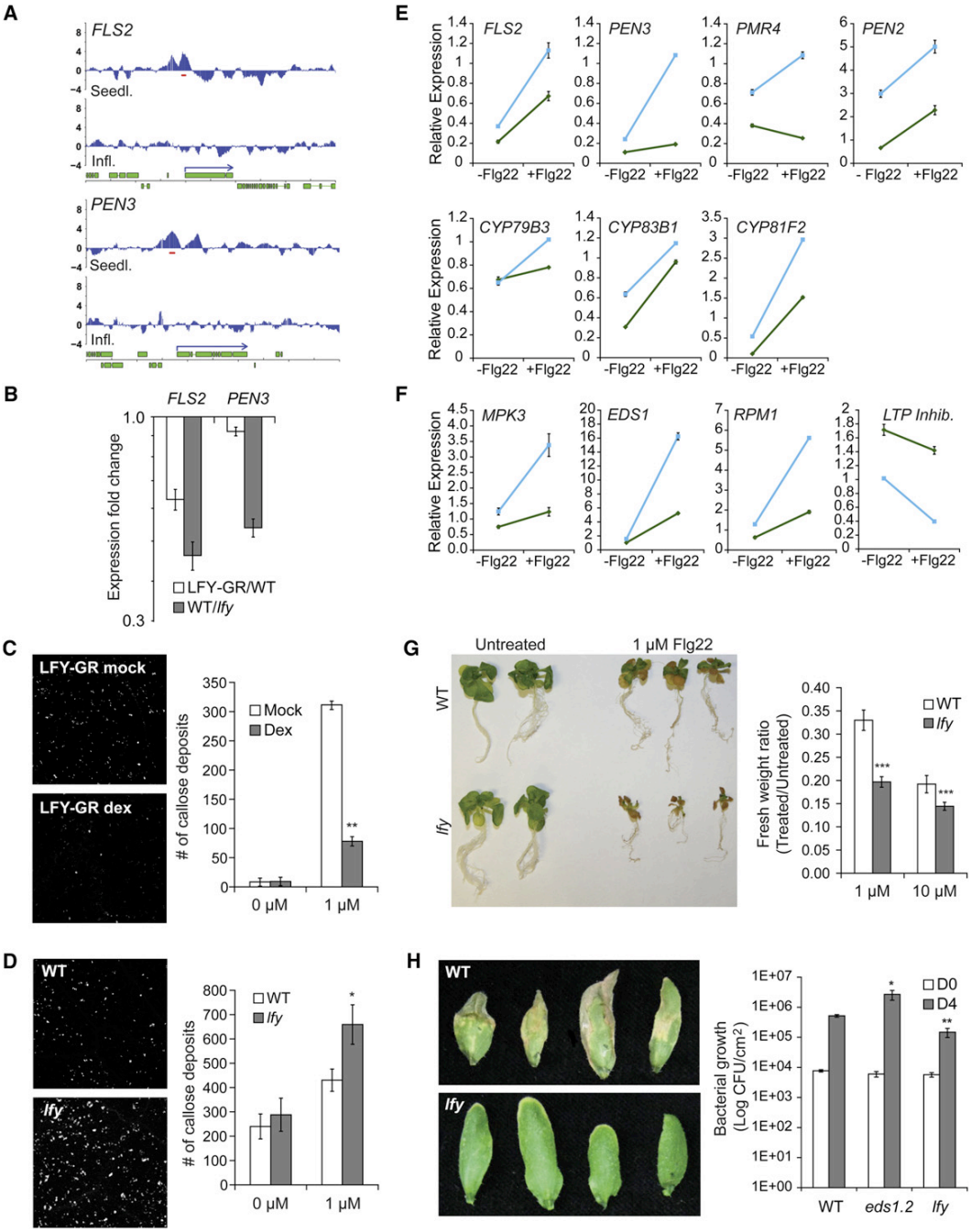


Figure 3. LFY Represses Responses to the Bacterial Flagellin Peptide flg22 and Pathogen Challenge

(A) Significant LFY binding to regulatory regions of *FLS2* and *PEN3* in seedlings (see Figure 1A for description of labels).

(B) Expression changes observed for *FLS2* and *PEN3* in wild-type (WT) and *lfy* mutant cauline leaves or after dexamethasone treatment of LFY-GR and wild-type seedlings. Seedling expression is based on our transcriptome analysis.

(C and D) Callose deposition triggered by flg22 in dexamethasone (dex) versus mock treated LFY-GR seedlings (C) and in *lfy* null mutant compared to wild-type cauline leaves (D). Right: quantification of callose foci from two independent experiments.

(E and F) Expression of direct LFY targets (*FLS2*, *PEN3*, *CYP79B3*, and *CYP83B1*) and other defense genes 1 hr after mock (-flg22) or flg22 infiltration of *lfy* (blue line) and wild-type (green line) cauline leaves. (E) Genes linked to flg22-induced callose deposition (Clay et al., 2009). (F) Flg22-regulated defense genes not linked to callose deposition.

(G) Growth suppression by flg22 in *lfy* mutant compared to wild-type seedlings. Right: Quantification of biomass. Left: Photograph after 8 days of treatment.

Developmental Cell
LEAFY Regulatory Targets

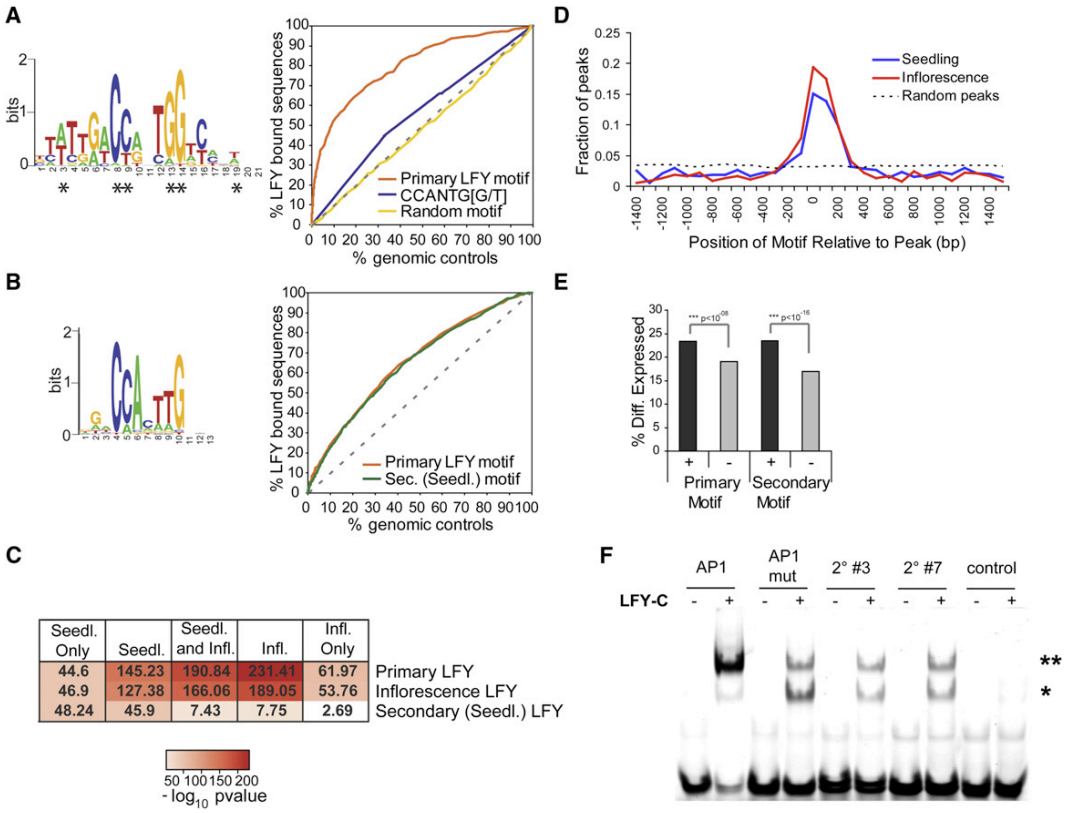


Figure 4. LFY Consensus Binding Motifs

(A) Left: primary LFY consensus motif identified by sequential motif analysis from a subset of the seedling-and-inflorescence bound regions. Asterisks: Nucleotides contacted by the LFY protein in LFY/DNA cocrystals (Hames et al., 2008). Right: Enrichment of the primary LFY motif, the previously known CCANTG[G/T] consensus (Busch et al., 1999), and a randomly permuted primary motif in all seedling-and-inflorescence bound regions based on receiver operating characteristic (ROC) curve analysis.

(B) Left: secondary LFY binding motif identified by sequential motif analysis from a subset of the seedling-only bound regions. Right: ROC curve analysis of both LFY motifs in all seedling-only bound sequences.

(C) Enrichment ($-\log_{10}$ p values) of de novo identified LFY motifs. Enrichment was tested in sequences bound by LFY in seedlings (Seedl.), in inflorescences (Infl.), at both stages (Seedl. and Infl.), in seedlings but not inflorescences (Seedl. Only) and in inflorescences but not seedlings (Infl. Only).

(D) Locations of the highest-scoring primary LFY motif within the 3000 bp surrounding LFY binding peak maxima (red and blue lines) or within 3000 bp of randomly generated peak maxima (dotted line).

(E) Presence of primary or secondary (seedling) LFY consensus motifs significantly enhances the probability of differential gene expression after LFY activation (logistic regression).

(F) Gel shift to test LFY binding to a primary motif (*AP1*), to an *AP1* motif containing only one LFY binding site (*AP1m*), to the secondary motif #3 (AT1G66480) and #7 (AT3G21890), and to an unrelated negative control motif. (**) dimeric LFY binding; (*) monomeric LFY binding.

See also Figure S5 and Table S3.

motif was identified when using a single motif prediction algorithm (Bailey and Elkan, 1995) (Figure S5C).

Both the primary and secondary motifs mapped close to the center of LFY binding peaks (Figure 4D; Figure S5D) and were present at regulatory regions of many known LFY target genes, as well as those identified here (Figures 1A and 2A; Figure S1E). Furthermore, the two LFY motifs together explain the majority of the LFY peaks observed (>72%; Table S3). Finally, the presence of a presumptive primary or secondary LFY motif near a given locus significantly enhanced the probability that it will be differentially expressed in response to LFY activation ($p < 10^{-08}$, logistic regression; Figure 4E).

(H) Right: Bacterial growth on adult rosette and cauline leaf discs of long-day grown plants after infection with *Pseudomonas syringae* pv. *tomato (Pst)* DC3000 measured at 0 and 4 days after inoculation. The *Ler eds1.2* mutant which exhibits enhanced susceptibility to bacterial pathogens was used as an infection control (Feys et al., 2005). Left: infiltrated cauline leaves at day 4.

(B and D–H): Data shown are mean ± SEM. Asterisks: Student's t test $p < 0.05$ (*), < 0.005 (**), < 0.0005(***). The same trend was observed in two independent experiments.

See also Figure S4.

Developmental Cell
LEAFY Regulatory Targets

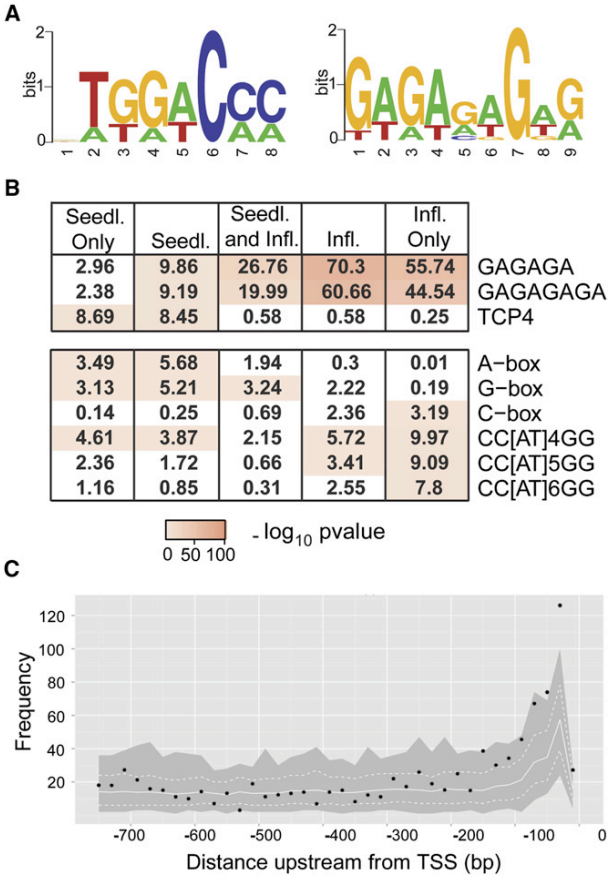


Figure 5. Identification of Potential LFY Cofactor Motifs

(A) De novo motif analysis of LFY-bound regions identified a motif with similarity to a class II TCP transcription factor binding motif (left), and a GA-rich motif (right).

(B) (Top) Enrichment ($-\log_{10}$ p values) of these cofactor motifs in LFY-bound regions. (Bottom) Enrichment of motifs of known LFY cofactors: bZIP (A-,G-, and C-boxes) and MADS ($CC[AT]_{4\text{-}6}GG$) (Krizek and Fletcher, 2005). No enrichment for homeodomain transcription factor binding sites was observed. Stages and categories are as described in Figure 4C.

(C) Black dots: Positional frequencies of GAGAGA repeats for LFY bound promoters in inflorescences. Gray ribbon: GAGAGA frequencies in TAIR9 promoters, with a solid white line indicating the mean and dashed lines indicating the 5th and 95th percentiles. The spike in GA-repeats at −35 bp from the TSS (position 0) in the genomic background shows an underlying tendency toward GA-rich sequences in core promoters of many *Arabidopsis* genes (Yamamoto et al., 2009).

To test whether the secondary motif can recruit LFY, we performed electrophoretic mobility shift assays (EMSAs) as well as yeast one-hybrid binding studies. The palindromic LFY binding site in *AP1* served as a representative primary motif (this study; Hames et al., 2008). The C-terminal DNA binding domain of LFY (LFY-C) bound to all seven tested secondary motifs based on EMSAs (Figure 4F; Figure S5E). The affinity of LFY-C for the secondary motifs was much lower than for the primary motif but comparable to that of an *AP1* motif in which one of the two LFY-bound half-sites (Hames et al., 2008) was mutated (Figure 4F; Figure S5E). In addition, LFY fused to the strong VP16 activation domain was able to confer increased growth of yeast to the fungal inhibitor aureobasidin A when recruited to a secondary motif, while LFY alone was not (Figure S5F), in agreement with prior studies which showed that LFY alone is not sufficient to activate transcription in the yeast one-hybrid assay (Parcy et al., 1998).

The de novo motif analysis also identified two potential LFY cofactor motifs, most notably a TGG(A/T)CC(C/A) motif and a GA-rich motif (Figure 5). The former is similar to the TCP4 transcription factor binding motif (Schommer et al., 2008). The TCP4 motif was significantly enriched in seedling-bound regions, while GA-repeat hexamer and octamer motifs were highly significantly enriched in inflorescence-bound regions (Figures 5B and 5C), suggesting that these elements may recruit stage-specific LFY cofactors. We also assessed the enrichment of known LFY cofactors (Figure 5B).

GA-repeat motifs are often found in Polycomb Responsive Elements (PREs) (Schuettengruber and Cavalli, 2009); hence, inflorescence-stage LFY targets may perhaps be repressed by polycomb group proteins at other developmental stages. Consistent with this hypothesis, the LFY inflorescence targets *AP3*, *AG*, and *SEP3* are regulated by polycomb repression (Goodrich et al., 1997; Liu et al., 2009b). In addition, our high-confidence inflorescence LFY target gene list was significantly enriched ($p < 0.05$) in genes repressed by polycomb-group proteins in seedlings (Table S4) based on queries of publicly available data sets (Oh et al., 2008; Turck et al., 2007) (http://affymetrix.arabidopsis.info/narrays/RefSearch.pl?ref_number=425).

DISCUSSION

Here, we present a genome-wide identification of direct LFY target genes. Many of the genes we identified are also bound by the known LFY cofactors SEP3 and AP1 (see http://published.genomics.upenn.edu/2010/LEAFY) (Irish, 2010; Liu et al., 2009b), in further support of their physiological relevance.

The three floral homeotic genes *AP3*, *AG*, and *PI* specify the identity of the reproductive organs of the flower, the stamens and carpels (Krizek and Fletcher, 2005). Regulation of the expression of these genes is, therefore, critical for reproductive success. Prior studies had revealed a direct role for LFY in induction of *AP3* and *AG* (Busch et al., 1999; Lamb et al., 2002). We show that LFY, in addition, directly upregulates the floral homeotic gene *PI*, in agreement with the previous demonstration that *PI* expression in developing flowers is strongly dependent on LFY (Weigel and Meyerowitz, 1993). We further report that LFY directly represses the polycomb group protein *EMF1* that prevents precocious activation of the floral homeotic genes (Calonje et al., 2008). In support of this, *emf1* null mutants are epistatic to *lfy* null mutants and LFY overexpression enhances a weak but not a null *emf1* mutant (Chen et al., 1997). Downregulation of *EMF1* by LFY may be required to overcome chromatin repression for initiation of flower patterning.

Finally, we show that LFY directly activates *SEP3* expression in the center of young flower primordia. LFY and SEP3 together induce *AP3*, *PI* and *AG* (Liu et al., 2009b). Thus, as reported for other developmental master regulators (for example, see

Tapscott, 2005), LFY activates expression of its own cofactor. The direct LFY target AP1 (Parcy et al., 1998; Wagner et al., 1999) also induces *SEP3* (Kaufmann et al., 2010; Liu et al., 2009b). *AP1* expression in flower primordia is redundantly activated by LFY, and by additional pathways such as the photoperiod flowering time pathway via FT and FD (Ruiz-Garcia et al., 1997; Wagner et al., 1999; Liu et al., 2009a). Hence, parallel converging pathways control *SEP3* induction (Figure 2D).

Our study, combined with previous findings (Calonje et al., 2008; Kaufmann et al., 2009, 2010; Liu et al., 2009b; Wagner et al., 1999), suggests that LFY operates as a highly connected regulatory 'hub' (Luscombe et al., 2004) upstream of three interlocking feed-forward loops that control the upregulation of *AP3*, *PI*, and *AG* expression. Such feed-forward loops function as signal persistence indicators and delay elements (Alon, 2007). In the context of floral homeotic gene regulation, the network that we describe would constrain upregulation of the floral homeotic genes to cells with robust accumulation of both LFY and SEP3, and ensure a delay in the induction of these genes relative to the time of floral initiation, preventing precocious differentiation and termination of the floral meristem.

How transcription factors regulate different target genes in different cell types and developmental stages is not well understood (Farnham, 2009). Our study uncovered two possible mechanisms that deserve further evaluation: selective LFY recruitment (by stage-specific *cis* and trans factors) and selective chromatin constraints (polycomb repression). Based on protein-binding microarrays, 50% of the transcription factors assayed recognize both a primary and a secondary consensus motif (Badis et al., 2009). We defined a palindromic primary LFY and a secondary (seedling) LFY consensus motif. LFY is predicted to bind the palindromic motif as a homodimer (Hames et al., 2008). LFY bound to the secondary motif in vitro; however the binding affinity is likely too low for the motif alone to recruit LFY in vivo, in particular at the seedling stage when less LFY protein is present. This finding, combined with the nonpalindromic nature of the secondary motif, suggests that LFY recruitment at the seedling stage may involve a second transcription factor. This factor might assist in recruitment by interacting with a nearby sequence and forming a heterodimeric complex with a LFY monomer (see Hollenhorst et al., 2009) or, alternatively, by modifying the affinity of the LFY homodimer for the secondary binding motif. Consistent with these ideas, LFY physically interacts with at least one other transcription factor (Liu et al., 2009b). The two LFY consensus motifs together explain the majority of the LFY binding peaks. The remaining binding events may be due to the presence of additional LFY motifs or, alternatively, to ''piggybacking'' of LFY to some regulatory regions by direct interaction with another transcription factor (Farnham, 2009).

We further identified motifs for potential LFY cofactors. Particularly intriguing among these were GAGA motifs preferentially enriched in LFY-bound sequences in inflorescences. This raises the possibility that, as in *Drosophila* and mammals (Schuettengruber and Cavalli, 2009; Sing et al., 2009), GAGA motifs may play a role in recruitment of polycomb group proteins in plants. Indeed, the high-confidence inflorescence LFY targets are enriched for genes whose expression is repressed by polycomb group proteins prior to flower formation. The identified direct repression of the polycomb regulator *EMF1* by LFY further suggests that LFY may play an active role in altering these chromatin constraints during flower patterning.

Developmental changes in resistance to pathogens and pests have been observed in many plant species (Develey-Riviere and Galiana, 2007; Herms and Mattson, 1992). Activation of defense responses can incur substantial fitness costs in terms of growth and reproduction (Heil, 2002; Tian et al., 2003). We report here that LFY modulates the plant immune response to pathogens. We show that LFY is required to repress plant responses to the bacterial MAMP flg22 and to reduce resistance to bacterial colonization and disease symptoms in leaves that form during the meristem identity transition. The data suggest that at this critical juncture in plant development LFY directs plant resources away from defense responses in these tissues and toward flower and fruit development in order to maximize reproductive fitness.

It remains to be seen whether a role for LFY in immune response is limited, for example, to plants with monocarpic life strategies like *Arabidopsis*, or observed more broadly. LFY orthologs have been identified in all species of land plants investigated, including nonflowering species (Moyroud et al., 2010). Our analysis of a public transcriptome data set (Maizel et al., 2005) revealed that LFY orthologs from additional flowering and nonflowering plant species also regulate target genes linked to plant defense (Figure S6). It is tempting to speculate that regulation of defense responses may be an ancestral LFY role; however, thus far there is no direct evidence for this conjecture and little is known about the molecular mechanisms underlying pathogen defense outside of seed plants.

Tradeoffs between stress avoidance and resource allocation to growth and reproduction are important for plant fitness and crop yield (Heil, 2002; Roux et al., 2006; Tian et al., 2003). Our studies suggest a possible mechanism for the coordination of developmental phase and defense programs. This finding is of potential ecological and also agricultural significance, given that many plant species of agricultural import including domesticated grains and many vegetable crops have monocarpic life strategies. A role for LFY in plant immune response may have gone unnoticed because pathogen response experiments are routinely performed on short-day grown plants that do not yet express *LFY*. It will be interesting to examine whether LFY links the onset of reproduction with additional, as yet undiscovered, stress responses.

EXPERIMENTAL PROCEDURES

Plant Materials and Growth Conditions

Plants of the Landsberg *erecta* (L*er*) accession were used. 35S:LFY-GR, *lfy-6* 35S:LFY-GR, and SEP3:SEP3-GFP were described previously (de Folter et al., 2007; Wagner et al., 1999). *ap1-1 cal-1* 35S:LFY-GR was a generous gift from Frank Wellmer. Because *lfy* mutants are sterile (Weigel et al., 1992), we obtained homozygous *lfy* null mutant seed by treating the parental *lfy-6* 35S:LFY-GR line with dexamethasone. Seeds cold-treated for 7 days at 4°C were grown in inductive conditions (continuous light or 16 hr long-day light conditions) or noninductive conditions (10 hr short-day light conditions) at 23°C at a fluence rate of 45 μmol/m^2 sec on 0.5× Murashige and Skoog plates (seedling experiments), or in soil (all other experiments).

Developmental Cell
LEAFY Regulatory Targets

RNA Analyses

Shoot apices from 9-day-old 35S:LFY-GR and L*er* seedlings were treated with 10 μM dexamethasone in 0.1% ethanol, for 4 hr as described (Wagner et al., 1999). RNA was extracted from four independent pools of apices using TRI-Reagent, column purified (RNeasy kit; QIAGEN), and amplified and labeled using NuGEN's Ovation RNA kits. Hybridization to the Affymetrix *Arabidopsis* ATH1 array was performed at the University of Pennsylvania Microarray Core Facility. Microarray data were processed using Bioconductor packages in R. Data were gcRMA normalized (Wu et al., 2004). Nonspecific filtering was performed with the MAS5.0 algorithm for genes that were "present" in at least two of four arrays in at least one treatment group (McClintick and Edenberg, 2006). Differentially expressed genes were identified using LIMMA (Smyth, 2004). For overlap analyses between LFY-bound genes and expression array data, only bound genes tested on the array were included, i.e., a probe set for the gene was printed on the array and passed our nonspecific filtering criteria.

For expression analysis of developmental regulators, 23-day-old long-day grown *ap1-1 cal-1* 35S:LFY-GR inflorescences were dipped for 1 min in a solution containing, 0.015% silwet77 and 0.01% ethanol alone or with 1 μM dexamethasone. RNA was isolated 2, 4, and 8 hr after treatment for dexamethasone-treated, and after 8 hr for mock-treated plants. For analysis of defense gene expression, the two basal-most fully expanded cauline leaves (long-day experiments) or fully expanded rosette leaves (short-day experiments) of L*er* and *lfy* null mutant plants were infiltrated with either 10 μM flg22 or water as previously described (Kim and Mackey, 2008). Leaves were harvested 1 and 3 hr after treatment. RNA isolation and cDNA synthesis were as described in (Yamaguchi et al., 2009). For all real-time PCR analyses the mean and standard error were determined using three technical replicates; one representative experiment is shown. Primers are listed in the Supplemental Experimental Procedures. In situ hybridization was performed as in (Yamaguchi et al., 2009) using probes previously described (Liu et al., 2009b).

ChIP-Chip Experiments

Chromatin immunoprecipitation (ChIP) was performed using 9-day-old seedlings treated with 10 μM dexamethasone, 0.1% ethanol, for 4 hr or 19-day-old untreated inflorescences with an anti-LFY antibody (Wagner et al., 1999) as described (Kwon et al., 2005) except that DNA was sonicated to an average size range of 300–500 bp.

ChIP and input DNA were amplified (see Supplemental Experimental Procedures) and hybridized to Affymetrix *Arabidopsis* 1.0R whole-genome tiling arrays. Three biological replicate 35S:LFY-GR IP samples and the corresponding input samples were hybridized for the seedling experiment while five biological replicate L*er* IP samples and the corresponding input samples were hybridized for the inflorescence experiment. The increased number of replicates enhanced peak detection for ChIP of endogenous LFY. Raw data were quantile normalized and significant binding regions were detected in CisGenome (Ji et al., 2008), employing the moving average method with a window size of 300 bp. Significant LFY binding peaks were assigned to genes using a custom Python script (see Supplemental Experimental Procedures for details).

Identification of High-Confidence LFY-Dependent and Coexpressed Genes

LFY-dependent genes were selected based on a statistically significant change in gene expression in *lfy* mutant relative to wild-type plants using LIMMA (FDR <0.05 and |fold change| >1.5; Smyth, 2004). Coexpressed genes were defined as those with expression patterns significantly correlated or anti-correlated with *LFY* or with the direct LFY targets *AP1*, *AP3*, or *AG* (bait genes). Known LFY target genes were included in the correlation analysis since target genes often exhibit a delay in gene expression relative to the transcription factor that regulates them (Chang et al., 2005). We used a Pearson's p value cutoff (<0.05) for the correlation analysis, which corresponds to an FDR of less than 0.15. See Supplemental Experimental Procedures for details regarding the expression data sets employed.

GO Term Analysis

Significant GO terms were identified using the GOstats Bioconductor package in R. Only GO terms annotated to more than ten genes were included. A combination of automated and manual curation was used to reduce redundancy of significant GO terms. Terms containing genes that overlapped by more than two-thirds were flagged and the more specific term was retained. In a few cases, the general term was deemed more informative and was retained instead.

De Novo Motif Prediction

Sequence regions of 500 and 750 base pairs surrounding the LFY peaks were used to generate sequence sets bound by LFY in (1) both seedlings and inflorescences, (2) seedlings but not inflorescences (seedling only), and (3) inflorescences but not seedlings (inflorescence only). For each sequence set and sequence length, three collections of 30 randomly pulled sequences from the top 50 most significantly LFY-bound sequences (FDR <0.01) were generated. The resulting 18 data sets were fed to a prediction pipeline consisting of five well cited prediction programs: MEME, AlignAce, MotifSampler, BioProspector, and Weeder (Bailey and Elkan, 1995; Hughes et al., 2000; Liu et al., 2001; Pavesi et al., 2001; Thijs et al., 2001). The most significantly enriched motifs from each of the five programs were aligned using a sliding window analysis for the shortest average Euclidean distance and merged additively, resulting in the primary, secondary (seedling only), and inflorescence only motifs. See Supplemental Experimental Procedures for further details.

Electrophoretic Mobility Shift Assay

The C-terminal DNA binding domain of LFY (LFY-C) was purified as described (Hames et al., 2008). For EMSAs, Cy5-dCTP labeled (GE Healthcare) oligos were used (see Supplemental Experimental Procedures). Binding reactions were performed in 20 μl binding buffer supplemented with 28 ng/μl fish sperm DNA (Roche), using 10 nM double-stranded DNA probe and 1 or 3 μM LFY-C. Binding reactions were loaded onto native 8% polyacrylamide gels and electrophoresed in 0.5 × TBE at 4°C. Gels were scanned on a Typhoon 9400 scanner.

Flg22 Treatment and Callose Staining

For seedling callose assays, 35S:LFY-GR and wild-type were grown for 9 days in long-day conditions in liquid culture essentially as previously described (Clay et al., 2009). Dexamethasone (10 μM) in 0.1% ethanol or 0.1% ethanol alone was added to the media, followed by 1 μM flg22 peptide (GenScript Corp, Piscataway, NJ) or water application 4 hr later. Plants were fixed after approximately 20 hr, washed, and stained with aniline blue as previously described (Clay et al., 2009). For leaf assays, plants were grown and infiltrated as for RNA analyses, except 1 μM flg22 was used. Leaves were fixed after 8 hr, washed, stained, and visualized as described above. Callose deposits were visualized on a Zeiss Axiovert microscope using UV illumination and a DAPI filter set. For growth inhibition in response to flg22, wild-type and *lfy* null mutant seedlings were treated as described (Gomez-Gomez et al., 1999). After 7 days of growth in long-day conditions seedlings were transferred to liquid culture. On day 11 seedlings were treated with 1 or 10 μM flg22 and photographed and weighed 8 days later.

Bacterial Growth Assays

Pseudomonas syringae pv. *tomato* (*Pst*) strain DC3000 was grown for 24 hr at 28°C on NYGA solid medium supplemented with 100 μg/mL rifampicin. Bolting plants (long-day experiments) or 4-week-old rosette leaves (short-day experiments) were spray-inoculated with bacterial suspensions at 4 × 10^8 cfu/ml in 10 mM $MgCl_2$ with 0.04% (v/v) Silwet L-77. In planta bacterial titers were determined 3 hr (day 0) and 4 days postinoculation by shaking leaf discs in 10 mM $MgCl_2$ with 0.01% (v/v) Silwet L-77 at 28°C for 1 hr as described previously (García et al., 2010; Tornero and Dangl, 2001; Vlot et al., 2008). Dilutions of the resulting bacterial suspension were then plated on NYGA solid medium containing rifampicin and grown at 28°C prior to colony counting. Titers were measured as the mean of four replicates (day 0) or six replicates (day 4), with each replicate containing three or more leaf discs. Bacterial numbers were compared between lines using a two-tailed Student's t test. The L*er eds1.2* mutant which exhibits enhanced susceptibility to *Pst* DC3000 (Feys et al., 2005) was used as an infection control.

Developmental Cell
LEAFY Regulatory Targets

ACCESSION NUMBERS

The raw data are deposited in NCBI's Gene Expression Omnibus and are accessible through the GEO Super Series accession number GSE28063. Processed data are available at our genome browser (http://published.genomics.upenn.edu/2010/LEAFY).

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures, four tables, and Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.devcel.2011.03.019.

ACKNOWLEDGMENTS

We thank Takashi Araki, Kim Gallager, Brian Gregory, Mary Mullins, Scott Poethig, John Wagner, and Matthew Willmann for comments on the manuscript, Daniel Simola for help implementing the ROC analysis, Frank Wellmer for the *ap1 cal* 35S:LFY-GR seed, Gerco Angenent for the SEP3:SEP3-GFP seed, and Hao Yu for the *SEP3* in situ probe. This work was funded by NSF IOS 0849298 to D.W., NIH Training grants T32HG000046 (Computational Biology) and T32-HD007516 (Developmental Biology) to C.M.W., funding from Agri-Food Canada, Agricultural Bioproducts Innovation Program to R.S.A., and a JSPS fellowship to A.Y.

Received: October 8, 2010
Revised: March 5, 2011
Accepted: March 29, 2011
Published: April 18, 2011

REFERENCES

Albani, M.C., and Coupland, G. (2010). Comparative analysis of flowering in annual and perennial plants. Curr. Top. Dev. Biol. *91*, 323–348.

Alon, U. (2007). Network motifs: theory and experimental approaches. Nat. Rev. Genet. *8*, 450–461.

Amasino, R. (2010). Seasonal and developmental timing of flowering. Plant J. *61*, 1001–1013.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. Science *324*, 1720–1723.

Bailey, T.L., and Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. Proceedings/international conference on intelligent systems for molecular biology. ISMB *3*, 21–29.

Blazquez, M.A., Soowal, L.N., Lee, I., and Weigel, D. (1997). LEAFY expression and flower initiation in Arabidopsis. Development *124*, 3835–3844.

Blazquez, M.A., Ferrandiz, C., Madueno, F., and Parcy, F. (2006). How floral meristems are built. Plant Mol. Biol. *60*, 855–870.

Bowman, J.L., and Floyd, S.K. (2008). Patterning and polarity in seed plant shoots. Annu. Rev. Plant Biol. *59*, 67–88.

Burow, M., Halkier, B.A., and Kliebenstein, D.J. (2010). Regulatory networks of glucosinolates shape Arabidopsis thaliana fitness. Curr. Opin. Plant Biol. *13*, 348–353.

Busch, M.A., Bomblies, K., and Weigel, D. (1999). Activation of a floral homeotic gene in Arabidopsis. Science *285*, 585–587.

Calonje, M., Sanchez, R., Chen, L., and Sung, Z.R. (2008). EMBRYONIC FLOWER1 participates in polycomb group-mediated AG gene silencing in Arabidopsis. Plant Cell *20*, 277–291.

Chang, W.C., Li, C.W., and Chen, B.S. (2005). Quantitative inference of dynamic regulatory pathways via microarray data. BMC Bioinformatics *6*, 44.

Chen, L., Cheng, J.C., Castle, L., and Sung, Z.R. (1997). EMF genes regulate Arabidopsis inflorescence development. Plant Cell *9*, 2011–2024.

Clay, N.K., Adio, A.M., Denoux, C., Jander, G., and Ausubel, F.M. (2009). Glucosinolate metabolites required for an Arabidopsis innate immune response. Science *323*, 95–101.

de Folter, S., Urbanus, S.L., van Zuijlen, L.G., Kaufmann, K., and Angenent, G.C. (2007). Tagging of MADS domain proteins for chromatin immunoprecipitation. BMC Plant Biol. *7*, 47.

Denoux, C., Galletti, R., Mammarella, N., Gopalan, S., Werck, D., De Lorenzo, G., Ferrari, S., Ausubel, F.M., and Dewdney, J. (2008). Activation of defense response pathways by OGs and Flg22 elicitors in Arabidopsis seedlings. Mol. Plant *1*, 423–445.

Develey-Riviere, M.P., and Galiana, E. (2007). Resistance to pathogens and host developmental stage: a multifaceted relationship within the plant kingdom. New Phytol. *175*, 405–416.

Dodds, P.N., and Rathjen, J.P. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. Nat. Rev. Genet. *11*, 539–548.

Farnham, P.J. (2009). Insights from genomic profiling of transcription factors. Nat. Rev. Genet. *10*, 605–616.

Feys, B.J., Wiermer, M., Bhat, R.A., Moisan, L.J., Medina-Escobar, N., Neu, C., Cabral, A., and Parker, J.E. (2005). Arabidopsis SENESCENCE-ASSOCIATED GENE101 Stabilizes and Signals within an ENHANCED DIEASE SUSCEPTIBILITY1 Complex in Plant Innate Immunity. Plant Cell *17*, 2601–2613.

García, A.V., Blanvillain-Baufumé, S., Huibers, R.P., Wiermer, M., Li, G., Gobbato, E., Rietz, S., and Parker, J.E. (2010). Balanced nuclear and cytoplasmic activities of EDS1 are required for a complete plant innate immune response. PLoS Pathog. *6*, e1000970.

Gomez-Gomez, L., Felix, G., and Boller, T. (1999). A single locus determines sensitivity to bacterial flagellin in Arabidopsis thaliana. Plant J. *18*, 277–284.

Goodrich, J., Puangsomlee, P., Martin, M., Long, D., Meyerowitz, E.M., and Coupland, G. (1997). A Polycomb-group gene regulates homeotic gene expression in Arabidopsis. Nature *385*, 44–51.

Guilfoyle, T.J., and Hagen, G. (2007). Auxin response factors. Curr. Opin. Plant Biol. *10*, 453–460.

Hames, C., Ptchelkine, D., Grimm, C., Thevenon, E., Moyroud, E., Gerard, F., Martiel, J.L., Benlloch, R., Parcy, F., and Muller, C.W. (2008). Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. EMBO J. *27*, 2628–2637.

Heil, M. (2002). Ecological costs of induced resistance. Curr. Opin. Plant Biol. *5*, 345–350.

Hempel, F.D., Weigel, D., Mandel, M.A., Ditta, G., Zambryski, P.C., Feldman, L.J., and Yanofsky, M.F. (1997). Floral determination and expression of floral regulatory genes in Arabidopsis. Development *124*, 3845–3853.

Herms, D.A., and Mattson, W.J. (1992). The dilemma of plants - to grow or defend. Q. Rev. Biol. *67*, 283–335.

Hill, T.A., Day, C.D., Zondlo, S.C., Thackeray, A.G., and Irish, V.F. (1998). Discrete spatial and temporal cis-acting elements regulate transcription of the Arabidopsis floral homeotic gene APETALA3. Development *125*, 1711–1721.

Hollenhorst, P.C., Chandler, K.J., Poulsen, R.L., Johnson, W.E., Speck, N.A., and Graves, B.J. (2009). DNA specificity determinants associate with distinct transcription factor functions. PLoS Genet. *5*, e1000778.

Honma, T., and Goto, K. (2000). The Arabidopsis floral homeotic gene PISTILLATA is regulated by discrete cis-elements responsive to induction and maintenance signals. Development *127*, 2021–2030.

Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J. Mol. Biol. *296*, 1205–1214.

Irish, V.F. (2010). The flowering of Arabidopsis flower development. Plant J. *61*, 1014–1028.

Jack, T. (2004). Molecular and genetic mechanisms of floral control. Plant Cell *16*, S1–S17.

Developmental Cell
LEAFY Regulatory Targets

Ji, H., Jiang, H., Ma, W., Johnson, D.S., Myers, R.M., and Wong, W.H. (2008). An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat. Biotechnol. *26*, 1293–1300.

Kaufmann, K., Muino, J.M., Jauregui, R., Airoldi, C.A., Smaczniak, C., Krajewski, P., and Angenent, G.C. (2009). Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the Arabidopsis flower. PLoS Biol. *7*, e1000090.

Kaufmann, K., Wellmer, F., Muino, J.M., Ferrier, T., Wuest, S.E., Kumar, V., Serrano-Mislata, A., Madueno, F., Krajewski, P., Meyerowitz, E.M., et al. (2010). Orchestration of floral initiation by APETALA1. Science *328*, 85–89.

Kim, D.H., Doyle, M.R., Sung, S., and Amasino, R.M. (2009). Vernalization: winter and the timing of flowering in plants. Annu. Rev. Cell Dev. Biol. *25*, 277–299.

Kim, M.G., and Mackey, D. (2008). Measuring cell-wall-based defenses and their effect on bacterial growth in Arabidopsis. Methods Mol. Biol. *415*, 443–452.

Kobayashi, Y., and Weigel, D. (2007). Move on up, it's time for change mobile signals controlling photoperiod-dependent flowering. Genes Dev. *21*, 2371–2384.

Komeda, Y. (2004). Genetic regulation of time to flower in Arabidopsis thaliana. Annu. Rev. Plant Biol. *55*, 521–535.

Krizek, B.A., and Fletcher, J.C. (2005). Molecular mechanisms of flower development: an armchair guide. Nat. Rev. Genet. *6*, 688–698.

Kwon, C.S., Chen, C., and Wagner, D. (2005). WUSCHEL is a primary target for transcriptional regulation by SPLAYED in dynamic control of stem cell fate in Arabidopsis. Genes Dev. *19*, 992–1003.

Lamb, R.S., Hill, T.A., Tan, Q.K., and Irish, V.F. (2002). Regulation of APETALA3 floral homeotic gene expression by meristem identity genes. Development *129*, 2079–2086.

Liu, C., Thong, Z., and Yu, H. (2009a). Coming into bloom: the specification of floral meristems. Development *136*, 3379–3391.

Liu, C., Xi, W., Shen, L., Tan, C., and Yu, H. (2009b). Regulation of flower patterning by flowering time genes. Dev. Cell *16*, 711–722.

Liu, X., Brutlag, D.L., and Liu, J.S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac. Symp. Biocomput. *2001*, 127–138.

Luscombe, N.M., Babu, M.M., Yu, H.Y., Snyder, M., Teichmann, S.A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. Nature *431*, 308–312.

Maizel, A., Busch, M.A., Tanahashi, T., Perkovic, J., Kato, M., Hasebe, M., and Weigel, D. (2005). The floral regulator LEAFY evolves by substitutions in the DNA binding domain. Science *308*, 260–263.

McClintick, J.N., and Edenberg, H.J. (2006). Effects of filtering by present call on analysis of microarray experiments. BMC Bioinformatics *7*, 49.

Michaels, S.D. (2009). Flowering time regulation produces much fruit. Curr. Opin. Plant Biol. *12*, 75–80.

Moyroud, E., Kusters, E., Monniaux, M., Koes, R., and Parcy, F. (2010). LEAFY blossoms. Trends Plant Sci. *15*, 346–352.

Mutasa-Gottgens, E., and Hedden, P. (2009). Gibberellin as a factor in floral regulatory networks. J. Exp. Bot. *60*, 1979–1989.

Nurmberg, P.L., Knox, K.A., Yun, B.W., Morris, P.C., Shafiei, R., Hudson, A., and Loake, G.J. (2007). The developmental selector AS1 is an evolutionarily conserved regulator of the plant immune response. Proc. Natl. Acad. Sci. USA *104*, 18795–18800.

Oh, S., Park, S., and van Nocker, S. (2008). Genic and global functions for Paf1C in chromatin modification and gene expression in Arabidopsis. PLoS Genet. *4*, e1000077.

Parcy, F. (2005). Flowering: a time for integration. Int. J. Dev. Biol. *49*, 585–593.

Parcy, F., Nilsson, O., Busch, M.A., Lee, I., and Weigel, D. (1998). A genetic framework for floral patterning. Nature *395*, 561–566.

Parcy, F., Bomblies, K., and Weigel, D. (2002). Interaction of LEAFY, AGAMOUS and TERMINAL FLOWER1 in maintaining floral meristem identity in Arabidopsis. Development *129*, 2519–2527.

Pavesi, G., Mauri, G., and Pesole, G. (2001). An algorithm for finding signals of unknown length in DNA sequences. Bioinformatics *17* (*Suppl 1*), S207–S214.

Poethig, R.S. (2003). Phase change and the regulation of developmental timing in plants. Science *301*, 334–336.

Roux, F., Touzet, P., Cuguen, J., and Le Corre, V. (2006). How to be early flowering: an evolutionary perspective. Trends Plant Sci. *11*, 375–381.

Ruiz-Garcia, L., Madueno, F., Wilkinson, M., Haughn, G., Salinas, J., and Martinez-Zapater, J.M. (1997). Different roles of flowering-time genes in the activation of floral initiation genes in Arabidopsis. Plant Cell *9*, 1921–1934.

Saddic, L.A., Huvermann, B., Bezhani, S., Su, Y., Winter, C.M., Kwon, C.S., Collum, R.P., and Wagner, D. (2006). The LEAFY target LMI1 is a meristem identity regulator and acts together with LEAFY to regulate expression of CAULIFLOWER. Development *133*, 1673–1682.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D., and Lohmann, J.U. (2005). A gene expression map of Arabidopsis thaliana development. Nat. Genet. *37*, 501–506.

Schmid, M., Uhlenhaut, N.H., Godard, F., Demar, M., Bressan, R., Weigel, D., and Lohmann, J.U. (2003). Dissection of floral induction pathways using global expression analysis. Development *130*, 6001–6012.

Schommer, C., Palatnik, J.F., Aggarwal, P., Chetelat, A., Cubas, P., Farmer, E.E., Nath, U., and Weigel, D. (2008). Control of jasmonate biosynthesis and senescence by miR319 targets. PLoS Biol. *6*, e230.

Schuettengruber, B., and Cavalli, G. (2009). Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. Development *136*, 3531–3542.

Sing, A., Pannell, D., Karaiskakis, A., Sturgeon, K., Djabali, M., Ellis, J., Lipshitz, H.D., and Cordes, S.P. (2009). A vertebrate Polycomb response element governs segmentation of the posterior hindbrain. Cell *138*, 885–897.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. *3*, Article3.

Steeves, T.A., and Sussex, I. (1989). Pattern in Plant Development (Cambridge, UK: Cambridge University Press).

Tapscott, S.J. (2005). The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription. Development *132*, 2685–2695.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics *17*, 1113–1122.

Tian, D., Traw, M.B., Chen, J.Q., Kreitman, M., and Bergelson, J. (2003). Fitness costs of R-gene-mediated resistance in Arabidopsis thaliana. Nature *423*, 74–77.

Tornero, P., and Dangl, J.L. (2001). A high-throughput method for quantifying growth of phytopathogenic bacteria in Arabidopsis thaliana. Plant J. *28*, 475–481.

Turck, F., Fornara, F., and Coupland, G. (2008). Regulation and identity of florigen: FLOWERING LOCUS T moves center stage. Annu. Rev. Plant Biol. *59*, 573–594.

Turck, F., Roudier, F., Farrona, S., Martin-Magniette, M.L., Guillaume, E., Buisine, N., Gagnot, S., Martienssen, R.A., Coupland, G., and Colot, V. (2007). Arabidopsis TFL2/LHP1 specifically associates with genes marked by trimethylation of histone H3 lysine 27. PLoS Genet. *3*, e86.

van Zanten, M., Snoek, L.B., Proveniers, M.C.G., and Peeters, A.J.M. (2009). The many functions of ERECTA. Trends Plant Sci. *14*, 214–218.

Vlot, A.C., Liu, P.P., Cameron, R.K., Park, S.W., Yang, Y., Kumar, D., Zhou, F.S., Padukkavidana, T., Gustafsson, C., Pichersky, E., et al. (2008). Identification of likely orthologs of tobacco salicylic acid-binding protein 2 and their role in systemic acquired resistance in Arabidopsis thaliana. Plant J. *56*, 445–456.

Wagner, D., Sablowski, R.W., and Meyerowitz, E.M. (1999). Transcriptional activation of APETALA1 by LEAFY. Science *285*, 582–584.

Weigel, D., and Meyerowitz, E.M. (1993). Activation of floral homeotic genes in Arabidopsis. Science *261*, 1723–1726.

Weigel, D., and Nilsson, O. (1995). A developmental switch sufficient for flower initiation in diverse plants. Nature *377*, 495–500.

Weigel, D., Alvarez, J., Smyth, D.R., Yanofsky, M.F., and Meyerowitz, E.M. (1992). LEAFY controls floral meristem identity in Arabidopsis. Cell *69*, 843–859.

Wellmer, F., Alves-Ferreira, M., Dubois, A., Riechmann, J.L., and Meyerowitz, E.M. (2006). Genome-wide analysis of gene expression during early Arabidopsis flower development. PLoS Genet. *2*, e117.

William, D.A., Su, Y., Smith, M.R., Lu, M., Baldwin, D.A., and Wagner, D. (2004). Genomic identification of direct target genes of LEAFY. Proc. Natl. Acad. Sci. USA *101*, 1775–1780.

Wu, Z., Irizarry, R.A., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004). A model based background adjustment for oligonucleotide expression arrays. J. Am. Stat. Assoc. *99*, 909–917.

Yamaguchi, A., Wu, M.F., Yang, L., Wu, G., Poethig, R.S., and Wagner, D. (2009). The microRNA-regulated SBP-Box transcription factor SPL3 is a direct upstream activator of LEAFY, FRUITFULL, and APETALA1. Dev. Cell *17*, 268–278.

Yamamoto, Y.Y., Yoshitsugu, T., Sakurai, T., Seki, M., Shinozaki, K., and Obokata, J. (2009). Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis. Plant J. *60*, 350–362.

Yant, L., Mathieu, J., and Schmid, M. (2009). Just say no: floral repressors help Arabidopsis bide the time. Curr. Opin. Plant Biol. *12*, 580–586.

Zipfel, C., Robatzek, S., Navarro, L., Oakeley, E.J., Jones, J.D., Felix, G., and Boller, T. (2004). Bacterial disease resistance in Arabidopsis through flagellin perception. Nature *428*, 764–767.

A link between LEAFY and B genes in *Welwitschia mirabilis* sheds light on ancestral mechanisms prefiguring floral development

Edwige Moyroud[a,b,c,d,e,1], Marie Monniaux[a,b,c,d], Emmanuel Thévenon[a,b,c,d], Renaud Dumas[a,b,c,d], Charles P. Scutt[e], Michael W. Frohlich[e,f,2] and François Parcy[a,b,c,d,2] [a] CEA, iRTSV, Laboratoire Physiologie Cellulaire et Végétale, F-38054 Grenoble, France. [b] CNRS, UMR5168, Grenoble, France.

[a] CEA, iRTSV, Laboratoire Physiologie Cellulaire et Végétale, F-38054 Grenoble, France. [b] CNRS, UMR5168, Grenoble, France. [c] Université Joseph Fourier-Grenoble I, UMR5168, Grenoble, France. [d], INRA, UMR1359, Grenoble, France. [e] Laboratoire de Reproduction et Développement des Plantes, UMR5667, CNRS, INRA, Université de Lyon, Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France. [f] Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3DS, UK. [1]Present address: Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK.

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Flowering plants (angiosperms) first appeared in the early Cretaceous, having evolved from a still not clearly identified gymnosperm ancestor. Major morphological differences and uncertain homologies between flowers and gymnosperm reproductive structures impede attempts to clarify flower origins based on fossils. Molecular developmental data may help elucidate the origin of important angiosperm features, such as the bisexuality of angiosperm flowers, derived from the typically unisexual reproductive structures of gymnosperms. Here, we use gene expression studies, coupled with state-of-the-art biophysical techniques, to infer likely properties of a gene regulatory network that controlled reproductive development in the last common ancestor of angiosperms and living gymnosperms. We show that *LEAFY(LFY)*-like genes in the gymnosperm *Welwitschia mirabilis* could regulate the expression of specific classes of MADS-box genes, as do their angiosperm counterparts during flower development, suggesting that these genes may already have been organized within a comparable pre-floral gene network before the appearance of flowers. *Welwitschia*, like most gymnosperms, contains two *LFY*-like genes, *WelLFY* and *WelNEEDLY (WelNDLY)*. We show that WelLFY shares, with its angiosperm ortholog, the biochemical capacity to regulate B-class MADS-box gene expression and we identify several *cis*-elements that contribute to this interaction. We also show that WelNDLY exhibits a distinct DNA binding specificity compared to that of WelLFY, which may have contributed to functional divergence of these factors.

gymnosperms | flower origin | bisexual structure | MADS-box | transcription factor

Introduction

One of the most important developmental changes in the evolutionary origin of the flower was the combination of male and female reproductive organs on a single growing axis (1-4). However, the origin of bisexuality in the angiosperms remains enigmatic (5-8). By comparing the genetic circuits that control the development of bisexual flowers and unisexual gymnosperm reproductive structures (GRS), we aim to reconstruct the developmental network that functioned in the last common ancestor of the living seed plants (angiosperms and extant gymnosperms). An understanding of this ancestral seed plant network should help to identify the subsequent molecular changes, which led to the appearance of the first flowers in the angiosperm lineage.

In angiosperms, male and female reproductive organ identity is controlled by the combinatorial expression of B- and C-class MADS-box genes: C gene expression confers female (carpel) identity in primordia arising from the center of the floral meristem, while combined B and C gene expression confers male (stamen) identity in primordia that form in the surrounding zone (9). Similarly in gymnosperms, C genes are expressed in both male and female GRS, while B genes are only expressed in male GRS (10-12). The expression of gymnosperm B- or C-class transgenes in flowering plants whose native B or C genes are inactivated by mutation is sufficient to restore near wild type flower development, suggesting that the biochemical properties of B and C homologs are widely conserved between seed plants (11, 12). To generate the bisexual structure of the first flowers, a C-class domain must have arisen next to a B+C-class domain on the same growing axis. Accordingly, several authors have proposed a change in the regulation of B and/or C genes as a crucial event on the lineage leading to the angiosperms (2, 9, 13, 14). In addition to B and C genes, gymnosperms also possess *AGL6* (15-17) and *Bsister* (18, 19) MADS-box lineages. It has been postulated that gymnosperm *AGL6*-like genes may be involved in the switch to reproductive development (15), while gymnosperm *Bsister* genes have been proposed to specifically regulate female developmental programs (18).

The *LEAFY*/*FLORICAULA* (*LFY*/*FLO*) gene encodes a unique plant transcription factor which, in angiosperms, patterns the floral meristem by regulating B- and C- class genes (20). In *Arabidopsis*, *LFY* is a direct activator of the B gene *APETALA3* (*AP3*) and the C gene *AGAMOUS* (*AG*) (21-24). All major groups of extant gymnosperms posses two paralogous *LFY*-like genes (1, 25), first identified in Monterey pine as *PRFLL* (26) and *NEEDLY* (*NLY*) (27), respectively, the only known exception being *Gnetum*, which possesses a single *LFY*-like gene (28). Phylogenetic analyses indicate that both *LFY* and *NLY* homologs were probably present in the last common ancestor of the living seed plants and that the *NLY* gene, retained in most gymnosperms, was subsequently lost in the angiosperm lineage before the radiation of the extant flowering plants (1). *LFY*-like genes are expressed in the developing GRS of all gymnosperms studied to date, consistent with a role for these genes in reproductive development (15, 26-31). A very complete study performed in three conifer species (25) also demonstrated that *LFY* and *NLY*

Reserved for Publication Footnotes

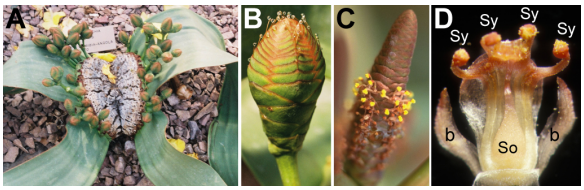


Fig. 1. – **Morphology of *Welwitschia mirabilis*** (A) *Welwitschia mirabilis* plant with female cones on stalks emerging from the scaly body at base of the leaves, at Huntington Botanical Gardens, San Marino, CA. (B) Female cone. (C) Male cone. (D) Close-up of a dissected male axillary unit showing the bracts (b) and the pollen-producing synangia (sy) at the tips of the antherophore which surround the sterile ovule (so) within its integument. One bract and part of the antherophore have been removed to show the central ovule.

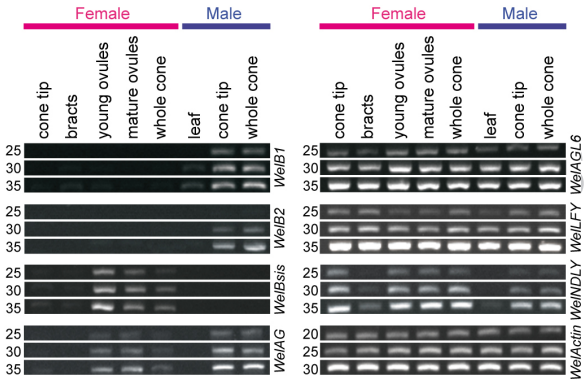


Fig. 2. – **Expression of *WelB1*, *WelB2*, *WelAG*, *WelBsister*, *WelAGL6*, *WelLFY* and *WelNDLY* in reproductive cones of *Welwitschia mirabilis*** Semi-quantitative RT-PCR profiles of *WelLFY*, *WelNDLY* and the *Welwitschia* MADS-box genes identified. The *Welwitschia* actin gene (*WelActin*) was used as a control. PCR cycles are given on the left. Names of the different tissues used are indicated above each lane.

exhibit divergent spatio-temporal expression patterns, suggesting these genes may make distinct contributions to GRS formation. However, it is not known whether gymnosperm *LFY*-like genes perform a similar function to their angiosperm counterparts by regulating B and C genes. A link between *LFY*-like genes and B/C genes in gymnosperms has been a central postulate of many hypotheses of flower origin (2, 9, 13, 14), though the existence of this link has never been demonstrated.

To elucidate the mechanisms that controlled reproductive development before the origin of the flower, we used a combination of gene expression and biophysical analyses to characterize the properties of a network involving LFY-like proteins and B- and C-class MADS-box genes in the gymnosperm *Welwitschia mirabilis*. *Welwitschia* is endemic to the costal desert of Namibia and Angola and presents numerous advantages as a gymnosperm for molecular-developmental studies: plants can make male reproductive structures (cones) in as little as 2 years from seed (32) and are small enough to be isolated in controlled environments. Although the plant body is famously bizarre, the reproductive structures are generalized; they have not lost numerous parts, as in their relatives *Gnetum* and *Ephedra*, nor do they have extensive fusions with the resulting morphological ambiguity of conifer cones. (33). *Welwitschia* cones show gradate development in which all stages are simultaneously available. The cones are borne on thin, branching stems that emerge from the "scaly body" between the leaves (Fig. 1A). Each cone bears a series of opposite (decussate) axillary fertile units subtended by sterile bracts (Fig.

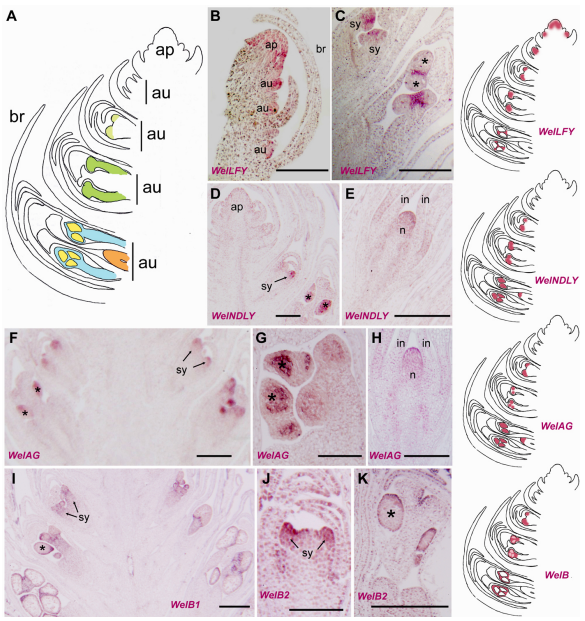


Fig. 3. – ***In situ* hybridization of *WelB1*, *WelB2*, *WelAG*, *WelLFY* and *WelNDLY* in male *Welwitschia* cones** (A) Schematic representation of a longitudinal section of a male cone showing the apex (ap), axillary units (au) and sterile bracts (br). The synangia primordia (light green) and young developing synangia (dark green) are visible in the second and third (from the top) axillary units. In later developmental stages, the pollen-producing tissues (yellow) differentiate from the rest of the synangia (blue). Inside this are the integuments of the sterile ovule (white) with the nucellus (orange) inside. (B-K) *In situ* hybridization of male cone sections using DIG-labelled RNA antisense *WelLFY* (B,C), *WelNDLY* (D,E), *WelAG* (F-H), *WelB1* (I) and *WelB2* (J,K) probes. The expression patterns are schematically summarized on the right. Scale bars = 200 µm, except for G scale bar = 100 µm. ap, apex; au, axillary unit; br, bract; sy, synangia; n, nucellus of the sterile ovule; in, integument of the ovule. Asterisks indicate the locations of sporangia.

Table 1.

		Length (bp)	$K_D^{App}1$ (µM)	$K_D^{App}2$ (µM)	χ^2
WelLFY-NC	*WelB2*	3377	235	0.045	2.6
	WelB1	2878	427	0.732	4.5
WelNDLY-NC	*WelB2*	3377	884	0.175	13.8
	WelB1	2878	2570	76.0	4.12

The sensogram curves corresponding to the association and dissociation between WelLFY-NC, WelNDLY-NC and the DNA molecules tested fitted best to the "heterogenous ligand model" that assumes the existence of two types of sites, consistent with the presence of a few high affinity sites ($K_D^{App}1$) among many low affinity sites ($K_D^{App}2$). If no high affinity sites are present then the model proposes two types of low affinity sites. bp, base pair; K_D^{App}, apparent dissociation constant, binding reaction was modelled with a heterogenous ligand model.

1B-C), such that newly formed units emerge at the cone tip and become progressively older towards the base. In female cones, each axillary unit comprises three pairs of opposite bracts surrounding a central fertile ovule. In male cones, each axillary unit (Fig. 1D) consists of two pairs of opposite bracts surrounding a tubular antherophore that bears six stalked synangia (the pollen producing organs). The antherophore encloses a sterile ovule

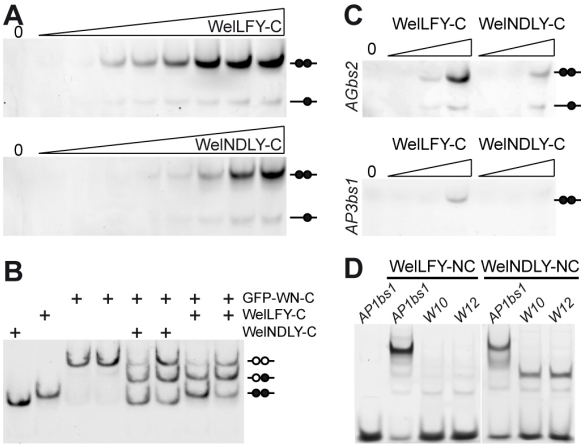


Fig. 4. – ***In vitro*** **comparison of WelLFY and WelNDLY DNA binding specificities** (A) Electrophoretic mobility shift assay (EMSA) with 10 nM fluorescent *AP1bs1* DNA and increasing concentrations of WelLFY-C or WelNDLY-C. Protein concentrations, from left to right, are 0, 0.025, 0.05, 0.1, 0.2, 0.4, 0.8, 1.5, 3 and 5 µM. The complexes corresponding to a monomer (one filled circle) or a dimer (two filled circles) bound to DNA are indicated on the right. (B) EMSA with 10 nM fluorescent *AP1bs1* DNA and various combinations of WelLFY-C (0.4 µM), WelNDLY-C (1.5 µM) and GFP-WelNDLY-C (1.5 µM). WelLFY-C or WelNDLY-C homodimer (two filled circles), GFP-WelNDLY-C homodimer (two open circles) and heterodimer involving a GFP and a non-GFP labelled protein (a filled and an open circle) are depicted. (C) EMSA with 10 nM fluorescent *AGbs2* or 10 nM *AP3bs1* and increasing concentrations of WelLFY-C or WelNDLY-C. Protein concentrations from left to right are 0, 0.1, 0.4 and 1.6 µM. (D) EMSA with 10 nM fluorescent *AP1bs1*, *W10* or *W12* DNA probe and 0 or 0.5 µMWelLFY-NC or 0.5 µMWelNDLY-NC.

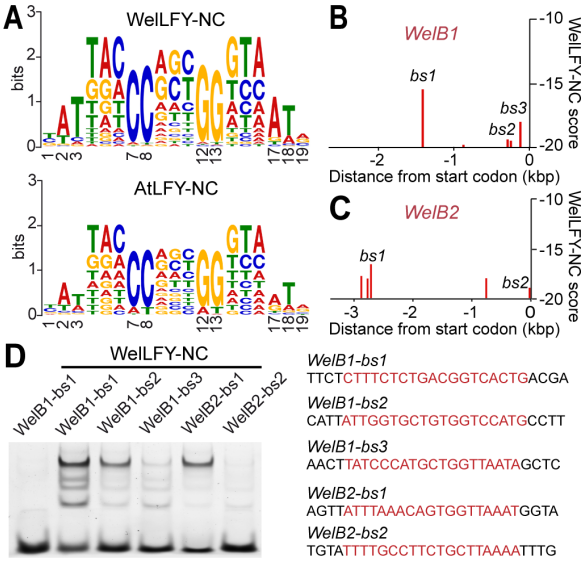


Fig. 5. – **Identification of WelLFY-NC binding sites upstream of the *WelB1* and *WelB2* coding sequences** (A) Logos of WelLFY-NC and AtLFY-NC Position Weight Matrix (B) Scores of the binding sites computed with the WelLFY-NC PWM along *WelB1* or (C) *WelB2* upstream sequences (bp, base pairs, counting from start codon). (D) EMSA with 10 nM fluorescent DNA corresponding to the five sites indicated in (C) and 0.5 µM of WelLFY-NC in lanes 2-6. The sequences of the five sites tested are indicated on the right, with the 19 bp motif indicated in red.

(Fig. 1D), which functions in the attraction of pollinators by producing sugar-containing droplets (34).

Orthologs of *LFY* and *NLY* in *Welwitschia* (*WelLFY* and *WelNDLY*) had previously been isolated (35), but not further characterized. Here, we isolated potential targets of *Welwitschia* LFY-like proteins, which corresponded to five distinct MADS-box genes of the B, C, AGL6 and B-sister clades and we showed that, among these genes, the expression of one C-class and two B-class genes were compatible with regulation by WelLFY and/or WelNDLY. Biophysical analyses of WelLFY revealed that this protein probably regulates B gene expression, as does its ortholog in angiosperms. These data constitute the first evidence at the biochemical level for a pre-floral network involving *LFY*-like and MADS-box genes in the last common ancestor of the living seed plants: a central postulate of many theories for the origin of flowers. Our analyses also reveal that WelLFY and WelNDLY show distinct DNA binding specificities and that only WelLFY is able to bind efficiently the promoters of the two B-class genes present in *Welwitschia*. Taken together, these observations point to two alternative scenarios for the roles of LFY and NLY in seed plant evolution, which should help to constrain present and future hypotheses for the origin of flowers.

Results

Expression of *Welwitschia* B- and C-class MADS-box genes parallels that of their angiosperm orthologs in male and female reproductive tissues

To identify potential targets of LFY-like proteins in *Welwitschia mirabilis,* we isolated five distinct MADS-box cDNAs from a *Welwitschia* male and female cone cDNA library. Subsequent phylogenetic analyses (Supp Fig.1) indicated these sequences to belong to the B (*WelB1* and *WelB2),* C *(WelAG),* AGL6 *(WelAGL6*) and Bsister (*WelBsister*) clades. We performed semi-quantitative RT-PCR analyses to elucidate the overall expression profiles of these genes in *Welwitschia* vegetative and reproductive tissues. These analyses (Fig.2) showed that *WelB1* and *WelB2* were both expressed exclusively in male cones, including the tip region where new axillary units are forming. The C gene, *WelAG*, was expressed in both male and female cones, but not in leaves or female bracts (Fig.2). *WelBsister* expression was detected exclusively in female *Welwitschia* cones (Fig.2), most strongly in ovules, though also in sterile bracts and in young axillary units at the cone tip. *WelAGL6* expression was detected in all tissues examined, including leaves (Fig.2). Finally, expression of *WelLFY* and *WelNDLY* was detected in all tissues examined, though *WelNDLY* was more weakly expressed in bracts and leaves than in the other tissues tested. Thus, the classes of *Welwitschia* MADS-box genes investigated showed expression profiles that parallel those of their orthologs in other gymnosperms and angiosperms. In particular, *Welwitschia* B and C genes were expressed specifically in reproductive tissues, with B gene expression further limited to male cones. As *LFY*-like genes are also expressed in both male and female reproductive organs of *Welwitschia*, the broad expression profiles of these genes are consistent with the possible regulation of B- and C-class MADS-box genes by *LFY*-like genes.

Expression of *Welwitschia LFY*-like genes precedes or parallels the activation of B- and C-class genes in male tissues

We performed *in situ* hybridizations to analyse the detailed spatiotemporal expression patterns of *LFY*-like genes and their potential targets during *Welwitschia* male cone development. Expression of *WelLFY* was first evident on the flanks of the cone apex where axillary units were about to form (Fig. 3A-B) and in very young bracts and emerging axillary unit primordia, just below the apical dome (Fig. 3B). *WelNDLY* transcripts, by contrast, were not detectable at the cone apex (Fig. 3D). As fertile units developed, the *WelLFY* signal disappeared from the bracts (Fig. 3B), but remained high in the synangium primordia (Fig. 3C) and later

became apparent in the upper part of the elongating synangia (Fig. 3C). As pollen-producing tissues started to differentiate, the *WelLFY* signal remained very strong, but became restricted to the regions bordering the pollen sacs. At this stage, *WelLFY* transcripts were clearly excluded from the tissues that would later generate pollen grains (indicated by asterisks in Fig. 3C). Similarly, *WelNDLY* was initially expressed in synangium primordia, though at slightly later stages this signal became restricted to the future pollen-producing tissues (indicated by asterisks in Fig. 3D). In addition, *WelNDLY* transcripts were observed at the top of the nucellus in the sterile ovule (Fig. 3E).

As with *LFY*-like genes, *Welwitschia* B (*WelB1* and *WelB2*) and C (*WelAG*) genes were also expressed during male axillary unit development. A strong *WelAG* signal first appeared when synangium primordia emerged (Fig. 3F). This signal was maintained as the primordia grew, though it became restricted to the cells that would give rise to pollen grains (Fig. 3F-G). *WelAG* expression was also visible at the top of the nucellus of the sterile ovule (Fig. 3H). The two B genes analyzed showed nearly identical expression patterns: both were first expressed throughout early male organ primordia, except in the centre of the axillary unit from which the ovule would emerge (Fig. 3I-J). These B gene signals were visible as a gradual coloration along the synangium stalks, reaching maximum intensity at the top of the pollen-producing organs. As sporogenous cells differentiated, B gene expression became restricted to the tissues enclosing these cells, similar to the expression of *WelLFY* (Fig. 3C, I and K). Based on these expression patterns, we conclude (i) that the expression of both *Welwitschia LFY-like* genes precedes or parallels the activation of B and C genes in male tissues, and (ii) that in later stages of male cone development, the expression domain of *WelLFY* and the two B genes becomes mutually exclusive from that of *WelNDLY* and *WelAG*. These data are consistent with a link between LFY/NDLY and the expression of B/C genes in *Welwitschia*. The striking correlations in expression patterns at later developmental stages suggest that, at least at these stages in male tissues, WelLFY may regulate B gene expression while WelNDLY may regulate C gene expression.

LFY orthologs in *Welwitschia* and *Arabidopsis* have conserved the DNA-binding specificity of their common ancestor, while that of WelNDLY differs

To investigate the roles of WelLFY and WelNDLY in transcriptional regulation, we characterized their DNA binding specificities. We produced recombinant versions of the DNA binding domains of these proteins (WelLFY-C, residues 247-411 and WelNDLY-C, residues 245-407) and analyzed their properties *in vitro*. Size-exclusion chromatography assays indicated that WelLFY-C and WelNDLY-C are monomeric in solution (Supp Table 1) and that both proteins can bind to *AP1bs1*: a DNA probe bearing the *Arabidopsis* LFY binding site from the *AP1* promoter (36). The binding profiles obtained for WelLFY-C and WelNDLY-C were reminiscent of *Arabidopsis* LFY-C (37) (Fig. 4A), suggesting that these *Welwitschia* factors, like LFY-C, bind to DNA as dimers. We confirmed this hypothesis by mixing WelNDLY-C with a GFP-tagged WelNDLY-C protein (Fig. 4B); a new complex of intermediate mobility formed, which corresponded to a WelNDLY-C/GFP-WelNDLY-C complex bound to *AP1bs1*. Mixing WelLFY-C and GFP-WelNDLY-C also gave rise to a novel, intermediate complex, demonstrating that WelLFY-C and WelNDLY-C can heterodimerize *in vitro* (Fig. 4B).

Our EMSA results indicate that WelLFY-C presents a higher affinity than WelNDLY-C for *AP1bs1* and for two other *Arabidopsis* LFY binding sites (*AGbs2* and *AP3bs1*, located in the regulatory second intron of *AG* and in the promoter of *AP3*, respectively), as complexes involving WelLFY-C are detected at lower protein concentrations (Fig. 4A, C). This suggests that, of the two *Welwitschia LFY*-like genes, WelLFY-C has the closest DNA preferences to *Arabidopsis* LFY. We then used a SELEX (Systematic Evolution of Ligands by EXponential enrichment) approach (38, 39), followed by massive sequencing, to characterize the DNA binding specificity of WelLFY and WelNDLY. For this, we used near full-length proteins (WelLFY-NC, WelNDLY-NC and AtLFY-NC for comparison) that included conserved N-terminal dimerization domains (40, 41). The logos of WelLFY-NC and AtLFY-NC show that the DNA binding preferences of these two factors are very similar (Fig. 4A). The analysis of sequences recovered using WelNDLY was less straightforward as it gave variable results, preventing us from generating a logo that faithfully reflects WelNDLY DNA binding specificity. Some sequences bound by WelNDLY clearly resemble those bound by LFY, whereas others are very different (Supp Fig. 2). For example, we could identify two sequences, W10 and W12, which are bound by WelNDLY-NC, but not by WelLFY-NC, to form a complex of lower mobility (Fig. 4D; Supp Fig. 2). These results indicate that WelLFY and WelNDLY have evolved distinct DNA binding behavior, along with their divergent expression patterns.

Biophysical evidence that WelLFY, like its angiosperm ortholog, regulates B gene expression

Based on gene expression patterns, we postulated that WelLFY may regulate *Welwitschia* B genes whereas WelNDLY may regulate *WelAG*. As our SELEX results provided a better characterization of DNA-binding specificity for WelLFY than for WelNDLY, we decided to investigate in detail the link between WelLFY and B- class genes. We isolated the sequences upstream of the *Welwitschia* B gene coding regions (2.8 kb for *WelB1* and 3.3 kb for *WelB2*) and tested their ability to interact with WelLFY, and also, for comparison, with WelNDLY. We used surface plasmon resonance (SPR) (42) to evaluate the specific interactions of each paralog with the upstream regions of the *Welwitschia* B genes, and with a 2.2 kb genomic fragment of the *WelTubulin* gene used as a negative control.

Neither WelLFY-NC nor WelNDLY-NC were able to bind efficiently to the *WelTubulin* upstream region (K_D^{App} best site ≈ 22 M) and the quality of the fits obtained for these interactions was low ($\chi^2 > 14$), as is often the case when only non-specific binding occurs (Table 1). However, the presence of sites of high affinity for WelLFY-NC (K_D^{App} ≈ 45 nM, χ^2 =2.6) was detected in the upstream sequence of *WelB2*. This DNA region could also contain binding sites for WelNDLY, but with a lower affinity (K_D^{App} ≈ 175 nM), and the insufficient quality of the fit (χ^2=13.8) prevented us from establishing the existence of such sites with confidence (Table 1). Results for *WelB1* were comparable: the presence of medium affinity binding sites was detected with WelLFY-NC (K_D^{App} ≈ 730 nM, χ^2= 4.5), whereas WelNDLY-NC was not able to bind efficiently to the upstream sequence of *WelB1* (K_D^{App} ≈ 76 µM, χ^2= 4.1). These results suggest that the upstream regions of both B genes in *Welwitschia* possess binding sites for WelLFY, but not for WelNDLY, consistent with the observed spatio-temporal correlation of WelLFY transcripts with those of B genes in male *Welwitschia* cones (Fig. 3). To identify those binding sites, we scanned the relevant upstream regions with our WelLFY-NC Position Weight Matrix (PWM) (Fig. 5A-C; Supp. Table 2). In each of the *WelB1* and *WelB2* upstream sequences that had been used in SPR analyses, we detected five motifs (Fig. 5B and C) predicted as high affinity WelLFY-NC binding sites (score > -20). We then used an EMSA assay to test whether WelLFY could bind efficiently to DNA probes corresponding to the three sites with the highest scores in the *WelB1* upstream sequence (named *WelB1-bs1, WelB1-bs2 and WelB1-bs3*) and the sites with, respectively, the highest (*WelB2-bs1*) and the lowest (*WelB2-bs2*) score above the -20 threshold in the *WelB2* upstream sequence (Fig. 5D). The results of these analyses showed that all sites tested except *WelB2-bs2* could interact efficiently with WelLFY-NC,

forming four distinct complexes (Fig. 5D). *WelB1-bs1* and *WelB2-bs1*, which were the motifs with the highest expected affinity according to WelLFY-NC PWM, gave the strongest interactions with WelLFY-NC, consistent with the predictions of our model (Fig. 5D). DNA/WelLFY-NC complexes could also be detected with *WelB1-bs2* and, to a lesser extent, with *WelB1-bs3*. In these cases, however, the prediction of our model was less accurate: WelLFY-NC PWM gave a higher score to *WelB1-bs3* than to *WelB1-bs2,* whereas in our EMSA assay, the intensity of the shift observed with *WelB1-bs2* was stronger than the signal detected with *WelB1-bs3* (Fig. 5D). Our results provide biochemical support for the proposition that WelLFY regulates the expression of both B genes, WelB1 and WelB2, in *Welwitschia*, corroborating our *in situ* hybridization data showing a near perfect coincidence of WelLFY and B gene expression in male cone tissues (Fig. 3). Thus, the positive regulation of B genes by LFY in a gymnosperm may constitute a significant element, conserved with angiosperms, of a pre-floral network that existed in the last common ancestor of the living seed plants.

Discussion

Evidence for a pre-floral gene network.

The conservation of B and C gene expression between angiosperms and gymnosperms, with C genes expressed in both sexes and B genes in male tissues only, was a first clue that similar genetic mechanisms may underlie reproductive development in all seed plants (9, 14, 43). As LFY regulates both B and C genes in angiosperms, it was proposed that *LFY*-like genes may also fulfil this role in gymnosperms. However, most gymnosperms contain two *LFY*-like genes of ancient origin, *LFY* and *NLY*, and previous expression studies did not provide conclusive evidence of separate or combined roles for these genes in the regulation of B and C genes (15, 16, 25, 26, 28-31).

In the present work, we began by showing that the presence and expression patterns of *LFY*-like and MADS-box B and C genes in *Welwitschia* resemble the situation in other gymnosperms: B and C genes show typical expression profiles in male and female reproductive tissues (Figs 2 and 3), while *WelLFY* and *WelNDLY* show somewhat broader expression patterns (Fig. 2). To provide firmer evidence than this correlation between *LFY*-like and MADS-box gene expression profiles, a genetic approach would be ideal. However, functional genetics in gymnosperms is not yet practicable, and so we used a range of state-of-the-art biochemical methods to demonstrate that an important part of the genetic network controlling flower development in *Arabidopsis* and other angiosperms could be conserved in *Welwitschia*. We used a combination of SELEX, SPR and EMSA analyses to show that WelLFY binds strongly and specifically to at least four sites in the presumptive promoters of two *Welwitschia* B genes and that WelLFY and its ortholog in *Arabidopsis* exhibit nearly identical DNA binding specificities. This conservation at the biochemical level, coupled with a detailed correlation between *LFY* and B gene expression in male *Welwitschia* reproductive tissues, combine to support the central tenet of hypotheses of flower origin: that *LFY*-like genes regulated the expression of specific classes of MADS-box genes in reproductive tissues before the appearance of the flower. In particular, the control of B class MADS-box genes by *LFY* homologs could predate the origin of the flower (Supp Fig. 3).

Other aspects of the pre-floral network in the common ancestor of the living seed plants can be tentatively inferred from our expression studies. For example, we observed that the expression of B and C genes overlaps as male organ primordia emerge at early developmental stages. However, B gene (*WelB1* and *WelB2)* expression stops once the pollen-producing organs are fully formed, while the C gene (*WelAG)* remains active throughout ovule development. This situation, also observed in *Gnetum* (44), substantiates the idea of a male program triggered by the combined expression of B and C genes, while the female program could be initiated by a C function alone. Since this feature is shared between extant gymnosperms and angiosperms, it is likely that the function of these genes was already established in the last common ancestor of extant seed plants (9, 11). The absence of *WelBsister* expression in the sterile ovules of male cones (Fig. 2) suggests *Bsister* genes may not be essential for ovule formation *per se,* but rather play a role in the later development of fully fertile ovules (e.g. in megaspore or megagametophyte formation). In *Arabidopsis,* the *Bsister* gene *TRANSPARENT TESTA16* is involved in seed coat pigmentation (45) and outer integument development (46, 47) so in both gymnosperms and angiosperms, *Bsister* genes appear to play late roles in female reproductive tissues: an observation which is not in agreement with the proposed ancestral role for these genes in specifying female organ identity (19). Expression analyses indicate that *WelLFY* is transcribed to high levels in vegetative tissues (Fig. 2), suggesting that *WelLFY* alone is not sufficient to confer reproductive fate. This situation is reminiscent of that in Norway spruce (*Picea abies*) and *Petunia*, in which *LFY*-like genes are expressed in both juvenile and reproductive tissues(15, 48). It has been proposed that the homolog of *AGL6* in Norway spruce could trigger reproductive fate (15), though our observations do not support a similar role in *Welwitschia* as *WelAGL6* transcripts are equally detected in all tissues examined (Fig. 2). In *Petunia*, the F-box protein DOUBLE TOP (DOT) is the limiting coregulator of LFY for flower specification (48) but homologs of DOT remain to be identified in gymnosperms.

The role of *LFY*-like genes in the evolution of the flower

Specialization of the roles of *LFY* and *NLY* homologs in gymnosperms has long been postulated but never firmly established (1, 13, 25). Here, we demonstrate that these two paralogs exhibit distinct DNA binding properties, which may allow them to fulfil different functions: LFY orthologs in angiosperms and gymnosperms have very similar DNA binding specificities (Fig. 5A), while the DNA-binding specificity of their paralog NLY has diverged (Fig. 4D). These observations explain why *LFY* from gymnosperms, expressed in *Arabidopsis* or tobacco (*Nicotiana tabacum*), rescued the *lfy* mutant phenotype more efficiently than *NLY* (27, 31, 41), and why in a yeast assay, WelNDLY failed to activate a reporter gene under the control of the LFY binding sites from *AP1* or *AG* regulatory regions (41) as efficiently as PRFLL (a homolog of *WelLFY* from Monterey pine). Our EMSA assays (Fig. 4D) also suggest that the complexes that WelNDLY forms with *W10* and *W12* differ from the DNA/protein complexes assembled with the canonical LFY binding sites (by having different protein:DNA ratios per complex or different quaternary structures). Taken together, these results indicate that WelNDLY evolved its own DNA binding specificity, and possibly an alternative mode of DNA binding from that of LFY and its orthologs.

A modification of the behavior of *LFY-like* genes and the homologs of floral homeotic regulators on the lineage leading to angiosperms has often been invoked (1, 2, 13). Our biochemical investigations point to two alternative scenarios, depending on whether *NLY* acquired its distinct properties before or after the divergence of gymnosperms and angiosperms (Supp Fig. 3). The first possibility is that NLY could have shared the same biochemical characteristics as LFY in the last common ancestor of extant seed plants and that both paralogs could have potentially regulated *B* gene expression in that ancestral species. Following the divergence of the extant gymnosperm and angiosperm lineages, *NLY* in gymnosperms could have evolved a different DNA binding specificity, along with a distinct expression pattern, allowing it to regulate the C gene. On the lineage leading to angiosperms, *NLY* could have failed to diverge and been lost,

while LFY became able to control C genes in addition to B genes. In this case, the loss of *NLY* in the angiosperm lineage could have no direct relationship to the origin of the flower. Alternatively (Supp Fig. 3), LFY and NLY could have already exhibited distinct properties in the last common ancestor of extant seed plants such that one gene (*LFY*) regulated *B*-genes, while the other (*NLY*) controlled *C*-gene expression. On the lineage leading to flowering plants, *LFY* could have gained the capacity to regulate C-gene expression while *NLY*, now redundant, was lost. Such a loss may not have occurred in gymnosperms as it may have been advantageous for gymnosperms to control B and C gene expression independently. In this case, the loss of *NLY* from the angiosperm lineage could have been an important corollary to the modifications that took place to the pre-floral network, causing the emergence of the first bisexual flowers (1).

Supporting Information

Methods

W. mirabilis Hook. f. tissues collected at the Huntington Botanical Gardens, San Marino, CA and at California State University, Fullerton CA were frozen in liquid nitrogen for mRNA extraction or fixed in 4% paraformaldehyde prior to *in situ* hybridization. *Welwitschia* cDNAs were isolated by a combination of RT-PCR and RACE-PCR, as described in Supplementary Methods. The phylogenetic placement of the fully sequenced cDNAs were assessed using ML phylogenetic reconstructions, as described in Supplementary Methods. The expression of LFY-like and MADS-box genes was initially analyzed in *Welwitschia* vegetative and reproductive tissues using semi-quantitative PCR, and then in detail during *Welwitschia* male cone development using *in situ* hybridization, as described in Supplementary Methods.

Recombinant C-terminal and near full-length versions of WelLFY and WelNDLY were expressed from pETM-11 (EMBL) and pET expression vectors and purified from *E. coli* cultures, and presumptive promoter regions upstream of *Welwitschia* MADS box genes were amplified from *Welwitschia* genomic DNA by anchored PCR, as described in Supplementary Methods. DNA binding behaviour of LFY-like proteins was analyzed using three complementary methods. Position Weight Matrices (PWM) describing LFY-like protein binding to all possible target sequences were derived using the SELEX procedure as previously described (39), while binding of LFY-like proteins to whole *Welwitschia* MADS-box gene promoters was quantified using Surface Plasmon resonance (SPR) (42). Individual candidate binding sites were identified by scanning the promoter sequences *in silico* using the PWMs and further tested together with known binding sites from *Arabidopsis* LFY target genes using EMSA assays, as fully described in Supplementary Methods.

Authors contribution

E.M., M.W.F. and F.P. designed research; E.M., M.W.F., M.M., R.D. and E.T. performed research; C.P.S. supervised the SPR experiments; E.M., M.W.F., M.M. and F.P. analyzed data; E.M., F.P., C.P.S. and M.W.F. wrote the paper. The authors declare no conflict of interest.

Acknowledgments.

We thank N. Maturen, K. James and K. Warner (NHM, London) for help with the *in situ* hybridization, M. Reymond (ENS Lyon) and A. Chaboud (IBCP Lyon) for their advice related to SPR experiments, N. Warthmann and D. Weigel for sequencing of the Selex samples and members of the Parcy laboratory for discussion. We thank J. Trager, J. Folsom (Huntington Botanical Gardens, San Marino, CA) and L. Song (California State University, Fullerton) for providing *Welwitschia* material, E. Meyerowitz for providing lab space for the processing of plant tissues, G. Theissen and R. Melzer for providing the MADS gene alignment from Becker and Theissen (2003), and B. Dentinger for help with phylogenetic analysis. This work was supported by funding from the CNRS (ATIP+ to F.P.), the ANR (Plant-TFcode to F.P. and C.P.S), PhD fellowships from the University J. Fourier, Grenoble (to E.M. and M.M.), the SYNTHESYS Project (to E.M.) and by the Floral Genome Project (NSF Plant Genome Research Program project DBI-0115684, M. W. F.) and by NSF DEB-9974374 (to M.W.F.).

1. Frohlich MW, and Parker,D.S. (2000) The mostly male theory of flower evolutionary origins: from genes to fossils. *Systematic Botany* 25:155-170.
2. Baum DA & Hileman LC (2006) A developmental genetic model for the origin of the flower. *In: Flowering and its manipulation—Ainsworth C, ed.*:Sheffield, UK: Blackwell Publishing. 3-27.
3. Theissen G & Melzer R (2007) Molecular mechanisms underlying origin and diversification of the angiosperm flower. *Ann. Bot. (London)* 100(3):603-619.
4. Rudall PJ & Bateman RM (2010) Defining the limits of flowers: the challenge of distinguishing between the evolutionary products of simple versus compound strobili. *Philos Trans R Soc Lond B Biol Sci* 365(1539):397-409.
5. Bateman RM, Hilton J, & Rudall PJ (2006) Morphological and molecular phylogenetic context of the angiosperms: contrasting the 'top-down' and 'bottom-up' approaches used to infer the likely characteristics of the first flowers. *J Exp Bot* 57(13):3471-3503.
6. Doyle JA (2008) Integrating molecular phylogenetic and paleobotanical evidence on the origin of the flower. *Int. J. Plant Sci.* 169:816-843.
7. Frohlich MW & Chase MW (2007) After a dozen years of progress the origin of angiosperms is still a great mystery. *Nature* 450(7173):1184-1189.
8. Mathews S & Kramer EM (2012) The evolution of reproductive structures in seed plants: a re-examination based on insights from developmental genetics. *New Phytologist* 194:910-923.
9. Becker A & Theissen G (2003) The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Mol Phylogenet Evol* 29(3):464-489.
10. Sundstrom J & Engstrom P (2002) Conifer reproductive development involves B-type MADS-box genes with distinct and different activities in male organ primordia. *Plant J* 31(2):161-169.
11. Winter KU, Saedler H, & Theissen G (2002) On the origin of class B floral homeotic genes: functional substitution and dominant inhibition in Arabidopsis by expression of an orthologue from the gymnosperm Gnetum. *Plant J* 31(4):457-475.
12. Zhang P, Tan HT, Pwee KH, & Kumar PP (2004) Conservation of class C function of floral organ development during 300 million years of evolution from gymnosperms to angiosperms. *Plant J* 37(4):566-577.
13. Albert VA, Oppenheimer DG, & Lindqvist C (2002) Pleiotropy, redundancy and the evolution of flowers. *Trends Plant Sci.* 7(7):297-301.
14. Theissen G & Becker A (2004) Gymnosperms orthologs of class B floral homeotic genes and their impact on understanding flower origin. *Critical Reviews in Plant Science* 23(2):129-148.
15. Carlsbecker A, Tandre K, Johanson U, Englund M, & Engstrom P (2004) The MADS-box gene DAL1 is a potential mediator of the juvenile-to-adult transition in Norway spruce (Picea abies). *Plant J* 40(4):546-557.
16. Mouradov A, *et al.* (1998) Family of MADS-Box genes expressed early in male and female reproductive structures of monterey pine. *Plant Physiol* 117(1):55-62.
17. Tandre K, Albert VA, Sundas A, & Engstrom P (1995) Conifer homologues to genes that control floral development in angiosperms. *Plant Mol Biol* 27(1):69-78.
18. Becker A, Bey M, Burglin TR, Saedler H, & Theissen G (2002) Ancestry and diversity of BEL1-like homeobox genes revealed by gymnosperm (Gnetum gnemon) homologs. *Dev Genes Evol* 212(9):452-457.
19. Becker A, *et al.* (2002) A novel MADS-box gene subfamily with a sister-group relationship to class B floral homeotic genes. *Mol Genet Genomics* 266(6):942-950.
20. Moyroud E, Kusters E, Monniaux M, Koes R, & Parcy F (2010) LEAFY blossoms. *Trends Plant Sci* .
21. Busch MA, Bomblies K, & Weigel D (1999) Activation of a floral homeotic gene in Arabidopsis. *Science* 285(5427):585-587.
22. Lamb RS, Hill TA, Tan QK, & Irish VF (2002) Regulation of APETALA3 floral homeotic gene expression by meristem identity genes. *Development* 129(9):2079-2086.
23. Lohmann JU, *et al.* (2001) A molecular link between stem cell regulation and floral patterning in Arabidopsis. *Cell* 105(6):793-803.
24. Parcy F, Nilsson O, Busch MA, Lee I, & Weigel D (1998) A genetic framework for floral patterning. *Nature* 395(6702):561-566.
25. Vazquez-Lobo A, *et al.* (2007) Characterization of the expression patterns of LEAFY/FLORICAULA and NEEDLY orthologs in female and male cones of the conifer genera Picea, Podocarpus, and Taxus: implications for current evo-devo hypotheses for gymnosperms. *Evol Dev* 9(5):446-459.
26. Mellerowicz EJ, Horgan K, Walden A, Coker A, & Walter C (1998) PRFLL--a Pinus radiata homologue of FLORICAULA and LEAFY is expressed in buds containing vegetative shoot and undifferentiated male cone primordia. *Planta* 206(4):619-629.
27. Mouradov A, *et al.* (1998) NEEDLY, a Pinus radiata ortholog of FLORICAULA/LEAFY genes, expressed in both reproductive and vegetative meristems. *Proc Natl Acad Sci U S A* 95(11):6537-6542.
28. Shindo S, Sakakibara K, Sano R, Ueda K, & Hasebe M (2001) Characterization of a FLORICAULA/LEAFY homologue of Gnetum parvifolium and its implications for the evolution of reproductive organs in seed plants. *J Plant Science* 162(6):1199-1209.
29. Dornelas MC & Rodriguez AP (2005) A FLORICAULA/LEAFY gene homolog is preferentially expressed in developing female cones of the tropical pine Pinus caribaea var. caribaea. *Genetics and Molecular Biology* 28(2):299-307.
30. Guo CL, Chen LG, He XH, Dai Z, & Yuan HY (2005) [Expressions of LEAFY homologous genes in different organs and stages of Ginkgo biloba]. *Yi Chuan* 27(2):241-244.
31. Shiokawa T, *et al.* (2008) Isolation and functional analysis of the CjNdly gene, a homolog in Cryptomeria japonica of FLORICAULA/LEAFY genes. *Tree Physiol* 28(1):21-28.
32. Van Jaarsveld E (1992) Welwitschia mirabilis in cultivation at Kirstenbosch. *Veld and Flora* 78:119-120.
33. Mundry M & Stützel T (2004) Morphogenesis of the reproductive shoots of Welwitschia mirabilis and Ephedra distachya (Gnetales), and its evolutionary implications *Organisms, Diversity & Evolution* 4:91-108.
34. Endress PK (1996) Structure and function of female and bisexual organ complexes in Gnetales. *International Journal of Plant SCiences* 157:S113-S125.
35. Frohlich MW & Meyerowitz EM (1997) The search for flower homeotic gene homologs in basal angiosperms and Gnetales: a potential new source of data on the evolutionary origin of flowers. *Int. J. Plant Sci.* 158(S131-S142).
36. Benlloch R, *et al.* (2011) Integrating long-day flowering signals: a LEAFY binding site is essential for proper photoperiodic activation of APETALA1. *Plant J* 67(6):1094-1102.
37. Hames C, *et al.* (2008) Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. *Embo J* 27(19):2628-2637.
38. Zhao Y, Granas D, & Stormo GD (2009) Inferring binding energies from selected binding sites. *PLoS Comput Biol* 5(12):e1000590.
39. Moyroud E, *et al.* (2011) Prediction of Regulatory Interactions from Genome Sequences Using a Biophysical Model for the Arabidopsis LEAFY Transcription Factor. *Plant Cell*

23(4):1293-1306.
40. Coen ES, *et al.* (1990) *Floricaula*: A homeotic gene required for flower development in Antirrhinum majus. *Cell* 63:1311-1322.
41. Maizel A, *et al.* (2005) The floral regulator LEAFY evolves by substitutions in the DNA binding domain. *Science* 308(5719):260-263.
42. Moyroud E, Reymond MC, Hames C, Parcy F, & Scutt CP (2009) The analysis of entire gene promoters by surface plasmon resonance. *Plant J* 59(5):851-858.
43. Tandre K, Svenson M, Svensson ME, & Engstrom P (1998) Conservation of gene structure and activity in the regulation of reproductive organ development of conifers and angiosperms. *Plant J* 15(5):615-623.
44. Becker A, Saedler H, & Theissen G (2003) Distinct MADS-box gene expression patterns in the reproductive cones of the gymnosperm Gnetum gnemon. *Dev Genes Evol* 213(11):567-572.
45. Nesi N, *et al.* (2002) The TRANSPARENT TESTA16 locus encodes the ARABIDOPSIS BSISTER MADS domain protein and is required for proper development and pigmentation of the seed coat. *Plant Cell* 14(10):2463-2479.
46. de Folter S, *et al.* (2006) A Bsister MADS-box gene involved in ovule and seed development in petunia and Arabidopsis. *Plant J* 47(6):934-946.
47. Prasad K, Zhang X, Tobon E, & Ambrose BA (2010) The Arabidopsis B-sister MADS-box protein, GORDITA, represses fruit growth and contributes to integument development. *Plant J* 62(2):203-214.
48. Souer E, *et al.* (2008) Patterning of inflorescences and flowers by the F-Box protein DOUBLE TOP and the LEAFY homolog ABERRANT LEAF AND FLOWER of petunia. *Plant Cell* 20(8):2033-2048.

CHAPITRE III :

Etude du changement de spécificité de LFY

chez *Physcomitrella patens*

Nos résultats de SELEX ont montré que la spécificité de reconnaissance de LFY était très bien conservée chez les plantes terrestres. Pourtant, chez la mousse *Physcomitrella patens*, le motif de liaison de PpLFY1 apparaît sensiblement différent. Ce résultat soulève plusieurs questions :

1) Le motif obtenu en SELEX est-il une représentation fidèle de la spécificité de PpLFY1 *in vitro* ? Il est important de vérifier ce motif expérimentalement puisque, pour les protéines NLY par exemple, le motif de SELEX n'a pas pu être validé.
2) Quelles sont les conséquences *in planta* de cette spécificité de liaison divergente ? La protéine PpLFY1 régule-t-elle un répertoire de gènes différents de celui régulé par LFY chez *Arabidopsis thaliana* ?
3) Comment expliquer, au niveau moléculaire et structural, un tel changement de spécificité alors que les 14 acides aminés impliqués dans le contact à l'ADN chez LFY sont identiques dans la séquence de PpLFY1 ?
4) Ce changement de spécificité a-t-il été brutal au cours de l'évolution, et si oui, comment a-t-il pu être toléré par les plantes ?

1) Optimisation de la matrice de PpLFY1 *in vitro*

a) Principe de la technique de QuMFRA

J'ai tout d'abord cherché à valider la spécificité de PpLFY1 obtenue par le SELEX par des mesures *in vitro*. Pour cela, j'ai utilisé la technique du QuMFRA (Quantitative Multiple Fluorescence Relative Affinity), qui consiste à mesurer, pour un oligonucléotide donné, son affinité relative pour la protéine par rapport à un oligonucléotide de référence (Man and Stormo, 2001). Chaque piste d'un gel retard contient un oligonucléotide test et un oligonucléotide référence, marqués avec des fluorophores différents, et la protéine va former des complexes de différentes intensités selon l'affinité qu'elle aura pour ces oligonucléotides (**Fig. 18A**). En mesurant l'intensité des retards observés, on peut facilement calculer le K_D relatif de la protéine pour l'oligonucléotide test (**Fig. 18B**).

La mesure de l'affinité relative de PpLFY1 pour un ensemble d'oligonucléotides nous permettra d'affiner la matrice obtenue par le SELEX pour que la corrélation entre les affinités prédites (via le score) et les affinités mesurées soit la meilleure possible.

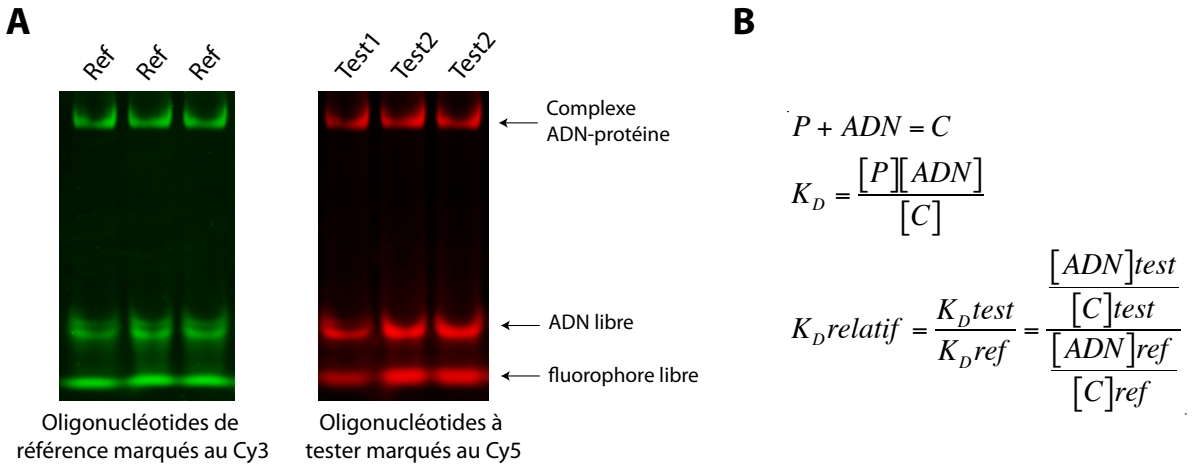


Figure 18: Principe du QuMFRA. **A:** Dans un même puits du gel retard, un oligonucléotide de référence (Ref) marqué au fluorophore Cy3 et un oligonucléotide test (Test1 ou Test2) marqué au fluorophore Cy5 sont mis en contact avec la protéine. Les fluorescences du Cy3 ou du Cy5 sont révélées séparément. **B:** Calcul de la constante de dissociation (K_D) relative d'un oligonucléotide pour la protéine. Le K_D relatif correspond au K_D de l'oligonucléotide test ($K_D$test), divisé par le K_D de l'oligonucléotide de référence ($K_D$ref). On s'affranchit ainsi de la concentration en protéine. En mesurant l'intensité de la bande d'ADN libre et de celle du complexe ADN-protéine pour les deux oligonucléotides, on en déduit le K_D relatif. P: protéine, C: complexe ADN-protéine, les concentrations sont figurées entre crochets.

b) Optimisation de la matrice de PpLFY1

La matrice de 19 pb, symétrique, obtenue en alignant les 2000 premières séquences du SELEX de PpLFY1, offre un pouvoir prédictif correct avec une corrélation entre les scores prédits et mesurés de 0,41 (**Fig. 19A**). Par analogie avec LFY-C, j'ai recherché dans le motif de liaison si certaines positions présentaient une dépendance entre elles, grâce au logiciel enoLOGOS (Workman et al., 2005). En effet, le calcul de scores suppose que chaque nucléotide d'une séquence participe de manière indépendante au score total. Pourtant, les positions d'un site de liaison ne sont pas toujours indépendantes entre elles, ce qui se traduit par un enrichissement significatif de certains doublets ou triplets de nucléotides (Man and Stormo, 2001). Le logiciel enoLOGOS calcule, à partir de l'alignement obtenu par MEME, la fréquence des différents couples de nucléotides à toutes les positions possibles, et renvoie un graphique représentant l'information mutuelle de chacun de ces couples (**Fig. 19B**). Nous avons constaté que les positions 2 et 18 d'une part, et 3 et 17 d'autre part, présentaient une dépendance entre elles. En calculant, à ces positions données, la fréquence des doublets de nucléotides et non pas des nucléotides indépendants, le calcul de scores prédit bien mieux la liaison de PpLFY1 *in vitro* (« matrice symétrique + doublets (1) », $r^2 = 0,63$) (**Fig. 19A**). Pour calculer les fréquences de chacun des doublets de nucléotides, un assez grand jeu de

séquences est nécessaire ; or seulement 514 séquences du SELEX (sur les 2000 données en entrée) sont en réalité alignées par MEME pour PpLFY1. J'ai donc calculé la fréquence des doublets sur la totalité des séquences uniques isolées par le SELEX (soit 49 379 séquences). Ce calcul a notablement amélioré la corrélation (« matrice symétrique + doublets (2) », r^2 = 0,76) (**Fig. 19A**), fournissant ainsi un modèle de bonne qualité pour prédire la liaison de PpLFY1 *in vitro*. L'ensemble de ces mesures a également confirmé que PpLFY1 possède une spécificité de liaison très différente de celle de LFYΔ, puisque les deux protéines reconnaissent des oligonucléotides totalement différents *in vitro* (**Fig. 19C**).

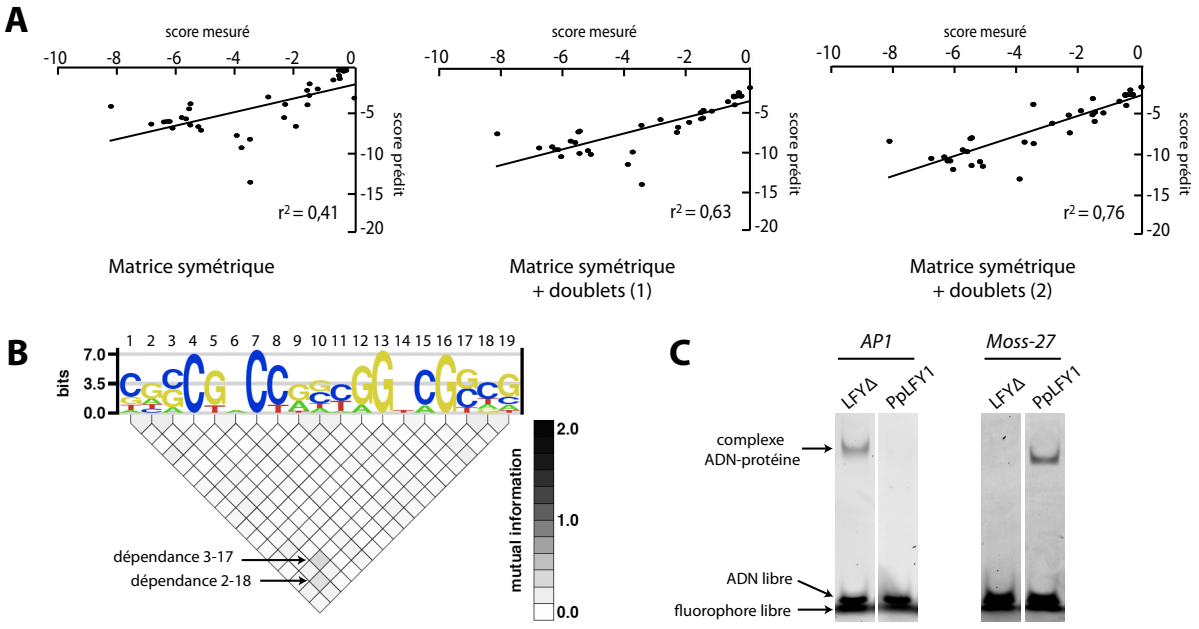


Figure 19 : Optimisation de la matrice poids/position de PpLFY1. A : Scores mesurés par QuMFRA et prédits par le modèle, et corrélation associée. Chaque point représente un oligonucléotide test mesuré par QuMFRA. La corrélation augmente du modèle de matrice symétrique (issu directement de l'alignement de MEME), au modèle « matrice symétrique + doublets (1) » qui intègre la fréquence des doublets aux positions 3-17 et 2-18 à partir des 514 séquences alignées par MEME. Dans le modèle « matrice symétrique + doublets (2) », la fréquence de ces doublets a été calculée à partir du lot total de 49 379 séquences uniques du SELEX de PpLFY1. **B :** Recherche de dépendances entre positions par le logiciel enoLOGOS, à partir de l'alignement proposé par MEME. enoLOGOS fournit un motif de liaison dont l'unité est l'information d'entropie relative, et un graphique représentant l'information mutuelle de tous les couples possibles de nucléotides. Deux couples de positions, indiqués par une flèche, ont une information mutuelle assez forte. **C :** Gel retard sur les oligonucléotides *AP1* et *Moss-27*, montrant que les spécificités des protéines LFYΔ et PpLFY1 sont radicalement différentes.

Par la suite, nous avons voulu déterminer les gènes cibles de PpLFY1, pour comprendre les conséquences du changement de spécificité observé sur la régulation des gènes cibles. Pour cela, j'ai initié deux approches : la prédiction bioinformatique des sites de liaison de PpLFY1 dans le génome de *P. patens,* et la recherche des gènes dérégulés *in planta* dans un contexte de surexpression de *PpLFY1*. Croiser ces deux approches nous permettra

d'obtenir une liste de gènes cibles directs potentiels pour PpLFY1 ; ce travail est encore en cours d'analyse, mais je vais présenter par la suite les résultats préliminaires que j'ai obtenus.

2) Prédiction des sites de liaison de PpLFY1 dans le génome de *P. patens*

Ayant développé un modèle prédictif des sites de liaison de PpLFY1, nous avons voulu identifier ces sites dans le génome de *P. patens*. J'ai calculé l'Occupation Prédite (POcc) de PpLFY1 pour toutes les séquences glissantes de 150 pb du génome de *P. patens*, et j'ai ensuite recherché quels gènes étaient les plus proches (par le promoteur ou le 3' UTR) des régions avec une POcc supérieure à 0,0001 : j'ai ainsi identifié 397 gènes associés à des régions favorables à la liaison de PpLFY1.

A

Scaffold	Position de la région	Pocc	Gène le plus proche de la région	Fonction de l'orthologue chez *A. thaliana*
scaffold_102	732496	0,134896	Pp1s102_114V6	
scaffold_83	1264270	0,121030	Pp1s83_183V6	deoxynucleoside kinase family
scaffold_157	444541	0,109921	Pp1s157_62V6	protein binding / zinc ion binding
scaffold_55	1324135	0,062770	Pp1s55_212V6	POL (poltergeist); protein serine/threonine phosphatase
scaffold_13	2461011	0,044338	Pp1s13_389V6	
scaffold_39	1183555	0,030136	Pp1s39_241V6	mRNA capping enzyme family protein
scaffold_33	450176	0,027046	Pp1s33_71V6	
scaffold_235	401876	0,024579	Pp1s235_86V6	haloacid dehalogenase-like hydrolase family protein
scaffold_43	793537	0,022949	Pp1s43_93V6	armadillo/beta-catenin repeat family protein / U-box domain-containing protein
scaffold_20	2407850	0,015167	Pp1s20_347V6	lipase family protein

B

Numéro du terme GO	Description du terme	Nombre de gènes associés à ce terme	Fréquence de ce terme dans le génome (f1, %)	Nombre de gènes avec POcc > 0,0001 associés à ce terme	Fréquence de ce terme dans les gènes avec POcc > 0,0001 (f2, %)	Enrichissement du terme (f2/f1)
GO:0000910	cytokinesis	25	0,052	5	0,241	**4,666**
GO:0015979	regulation of photosynthesis, dark reaction	50	0,103	5	0,241	**2,333**
GO:0007275	multicellular organismal development	35	0,072	4	0,193	**2,666**
GO:0018022	eukaryotic translation initiation factor 4F complex	10	0,021	3	0,145	**6,998**
GO:0009934	regulation of meristem structural organization	12	0,025	3	0,145	**5,832**
GO:0048573	photoperiodism, flowering	20	0,041	3	0,145	**3,499**
GO:0048481	ovule development	25	0,052	3	0,145	**2,799**
GO:0051726	G1 phase of mitotic cell cycle	27	0,056	3	0,145	**2,592**
GO:0016740	cell wall mannoprotein biosynthetic process	30	0,062	3	0,145	**2,333**

Figure 20 : Bilan des meilleures régions de 150 pb de liaison de PpLFY1 prédites dans le génome de *P. patens*. A : 10 meilleures régions identifiées, gènes associés et leurs fonctions présumées grâce à celle de leur orthologue chez *A. thaliana*. **B :** Enrichissement de quelques termes de Gene Ontology dans les gènes associés aux régions à la POcc > 0,0001. On considère qu'il y a enrichissement si f2/f1 est supérieur à 2.

La fonction de nombreux gènes de *P. patens* n'est pas encore établie ; j'ai donc recherché la fonction de leur orthologue (lorsqu'il en existe un) chez *A. thaliana* pour m'approcher de données fonctionnelles. Les dix premiers sites identifiés, par exemple, sont associés à des gènes reliés à des fonctions cellulaires très variées (**Fig. 20A**). J'ai recherché ensuite les termes de Gene Ontology (GO) auxquels ces gènes étaient associés, pour

déterminer si certains termes étaient particulièrement présents chez les gènes prédits (**Fig. 20 B**). On retrouve, parmi de nombreux autres termes, un enrichissement pour des termes reliés à la division cellulaire (cytodiérèse, phase G1), l'organisation méristématique, ou encore les remodellages de la paroi cellulaire, ce qui est cohérent avec la fonction présumée (via le phénotype mutant) de PpLFY1 sur le contrôle de la division cellulaire. Ces analyses restent très préliminaires puisque les sites de liaison sont uniquement prédits et non validés expérimentalement, et que la fonction des gènes associés aux bons sites de liaison n'est que supposée grâce à la fonction de l'orthologue présumé chez *A. thaliana*. De plus, nous savons que pour LFY-C, au mieux 25% des gènes prédits sont liés par la protéine *in vivo*. Croiser ces prédictions avec l'approche *in planta* est donc essentiel.

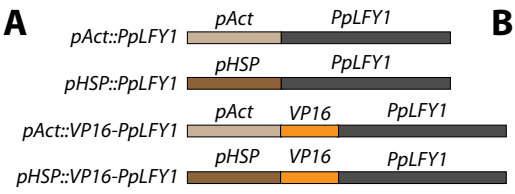
3) Recherche des gènes régulés par PpLFY1

a) Obtention de plantes surexprimant *PpLFY1*

Pour rechercher les gènes régulés par PpLFY1 *in planta*, j'ai mis en place la culture de *P. patens* au laboratoire, avec la collaboration de Fabien Nogué et Florence Charlot (Institut Jean-Pierre Bourgin, INRA Versailles). Nous avons ensuite choisi de générer des plantes surexprimant *PpLFY1*. En effet, le double mutant *pplfy1 pplfy2* arrête son développement après formation du zygote et ne produit pas de sporophyte, tissu principal où sont exprimés *PpLFY1/2* (Tanahashi et al., 2005) ; la recherche des gènes cibles de PpLFY1/2 n'est donc pas possible chez le double mutant. L'effet de la surexpression de *PpLFY1* restait par contre inconnu jusqu'à présent.

P. patens se développe en tant qu'espèce modèle, de par sa position phylogénétique et son cycle de vie assez rapide, mais également grâce à l'efficacité de son système de recombinaison homologue qui permet de générer facilement des plantes transgéniques en contrôlant le lieu d'insertion des transgènes (Schaefer, 2001; Prigge and Bezanilla, 2010). Pour surexprimer *PpLFY1*, j'ai créé des constructions constituées du promoteur *pAct* (promoteur du gène de l'Actine chez le riz) ou du promoteur *pHSP* (promoteur inductible par choc thermique) (Saidi et al., 2005), contrôlant l'expression de *PpLFY1* ou d'une fusion *VP16-PpLFY1* (fusion d'un domaine activateur de la protéine VP16 à PpLFY1) (**Fig. 21A**). Cette fusion a été créée puisque LFY seul ne peut pas activer la transcription génique chez la levure, sauf s'il est fusionné à un domaine activateur (tel que celui de la protéine VP16) (Parcy, 2005) ; il est donc possible que la surexpression de *PpLFY1* seule ne dérégule que très

peu de gènes. Ces constructions ont ensuite été insérées dans le génome de *P. patens* par recombinaison homologue, dans une région qui autorise une forte expression génique. Les constructions contiennent un gène de résistance à un antibiotique, permettant deux étapes successives de sélection des clones transformés ; après une étape de sélection, on isole des transformants stables (ayant intégré la construction dans leur génome) et instables (pas d'intégration génomique) que l'on appelle des transformants primaires. Après une deuxième étape de sélection, on considère que seuls les transformants stables (donc secondaires) sont isolés.



B

Construction	Nombre de clones résistants			
	1ère sélection (transformants stables et instables)	2ème sélection (transformants stables)		
		Nombre total de clones	Nombre de clones génotypés	Nombre de clones possédant l'insertion recherchée
pHSP::PpLFY1	1200	55	11	11
pAct::PpLFY	1500	59	15	6
pHSP::VP16-PpLFY1	110	14	3	3
pAct::VP16-PpLFY1	35	7	1	0

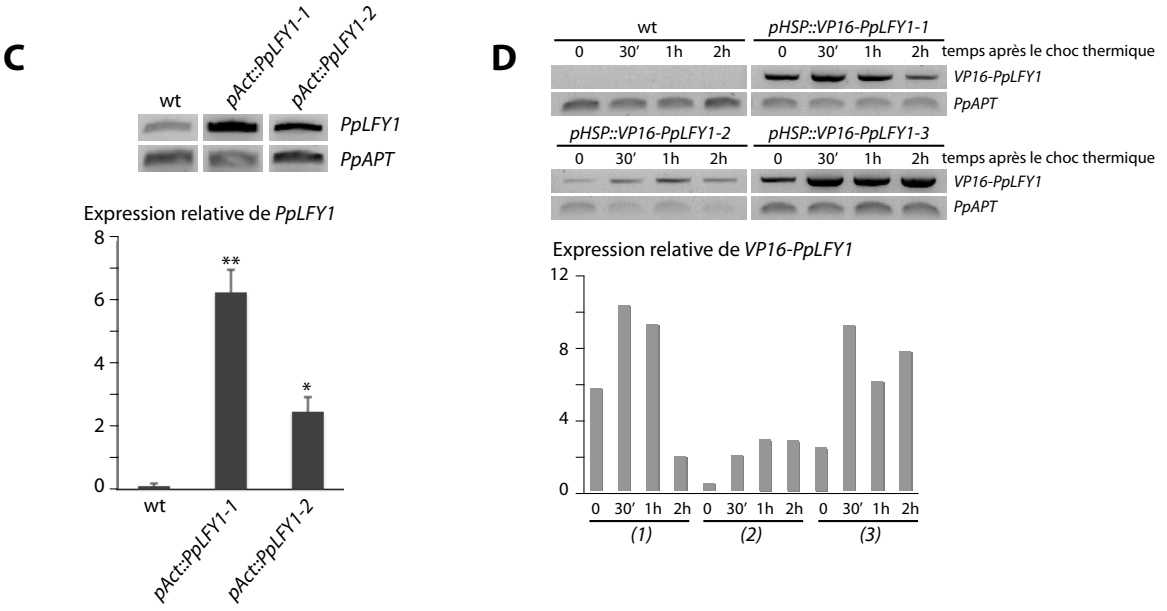


Figure 21 : Bilan des constructions transformées chez *P. patens*, et analyse des niveaux d'expression de *PpLFY1* chez certains clones. A : Organisation des constructions utilisées pour surexprimer *PpLFY1* ou *VP16-PpLFY1* chez *P. patens*. De part et d'autre de chaque construction représentée se trouve une zone d'homologie pour un locus génomique. **B :** Nombre de transformants isolés pour chacune des constructions de A, après une ou deux étapes de sélection sur antibiotique (voir Matériel et Méthodes). Les constructions comportant la fusion *VP16-PpLFY1* génèrent un nombre de transformants très réduit. **C :** Niveaux d'expression relatifs de *PpLFY1* chez deux clones de type *pAct::PpLFY1* analysés par RT-PCR. **D :** Niveaux d'expression relatifs de *VP16-PpLFY1* chez trois clones de type *pHSP::VP16-PpLFY1* analysés par RT-PCR, avant (0) ou à différents temps (30 minutes, 1h ou 2h) après 1h de choc thermique à 37°C. Les niveaux d'expression relatifs sont calculés par rapport au gène de référence *PpAPT* (Trouiller et al., 2006).

Les constructions comprenant la fusion VP16-PpLFY1 n'ont généré que très peu de transformants primaires, et aucun transformant secondaire *pAct::VP16-PpLFY1* n'a pu être validé, contre seulement 3 pour la construction *pHSP::VP16-PpLFY1* (**Fig. 21B**). La fusion

VP16-PpLFY1 semble donc létale pour la plante, alors que la transformation d'une construction *pAct::VP16* seule génère de nombreux transformants qui se développent sans défauts phénotypiques apparents (**Fig. 22O**). Les constructions *pAct::PpLFY1* et *pHSP::PpLFY1* ont toutes les deux généré de nombreux clones. Pour *pAct::PpLFY1*, après validation de l'insertion génomique, le tissu gamétophytique d'une dizaine de clones a été analysé par RT-PCR pour mesurer le niveau d'expression de *PpLFY1*. De manière surprenante, la plupart des clones montraient un faible niveau de surexpression de *PpLFY1* (2 à 3 fois plus que chez la plante sauvage). J'ai néanmoins pu isoler deux clones (*pAct::PpLFY1-1* et *pAct::PpLFY1-2*) montrant une forte surexpression de *PpLFY1* (respectivement 45 et 32 fois plus que chez la plante sauvage) (**Fig. 21C**). Les 3 clones *pHSP::VP16-PpLFY1* ont également été analysés par RT-PCR, avant ou après choc thermique à 37°C ; l'induction de l'expression de *VP16-PpLFY1* a été confirmée pour ces 3 clones, avec des cinétiques et des intensités d'induction différentes (**Fig. 21D**). Toutes les surexpressions ont été confirmées au niveau protéique par Western-Blot.

b) Analyse phénotypique des plantes surexprimant *PpLFY1*

P. patens se développe en deux phases : gamétophytique (haploïde) puis sporophytique (diploïde). Au stade gamétophytique, la plante produit tout d'abord des filaments de cellules chlorophylliennes (protonémas ou caulonémas), qui vont ensuite construire des tiges feuillées (les gamétophores). Ces gamétophores portent à leur sommet les structures reproductrices mâles ou femelles. Après fécondation, le sporophyte se développe à partir des structures reproductrices femelles, au sommet du gamétophore, et produit les spores qui vont rétablir l'haploïdie et germer en filaments (Cove, 2005).

La plante *pAct::PpLFY1-1* montre une sénescence précoce des filaments au stade gamétophytique, et les caulonémas formés à la périphérie du clone sont courts et peu ramifiés (**Fig. 22 A-D**). Des gamétophores se développent correctement, mais ils sont d'aspects plus courts et trapus que ceux de la plante sauvage (**Fig. 22 E-G**). L'analyse de ce clone au stade sporophytique est en cours. La plante *pAct::PpLFY1-2* semble montrer également des caulonémas et des gamétophores plus courts que chez la plante sauvage (non présenté dans la figure), mais ce phénotype apparaît moins marqué que pour le clone *pAct::PpLFY1-1*, en accord avec les niveaux de surexpression de *PpLFY1* de ces deux clones (**Fig. 21C**). La surexpression de *PpLFY1* a donc des effets phénotypiques légers mais visibles, et semble de manière générale affecter la croissance de la plante. Les plantes *pHSP::VP16-PpLFY1* ont

quant à elles un phénotype très marqué. Les clones 1 et 3, sans choc thermique, présentent déjà une sénescence avancée des filaments, alors que le clone 2 ne semble pas affecté (**Fig. 22 H-K**), à nouveau en accord avec les niveaux d'expression de *VP16-PpLFY1* déterminés par RT-PCR (**Fig. 21D**). Après 3 jours de choc thermique (à raison d'1h à 37°C par jour), les clones 1 et 3 ne montrent que très peu de filaments encore vivants, et le clone 2 présente une sénescence avancée des filaments (**Fig. 22 L-N**). Cet effet drastique n'est pas dû au domaine VP16 seul, puisque les plantes *pAct::VP16* ont un phénotype sauvage (**Fig. 22O**).

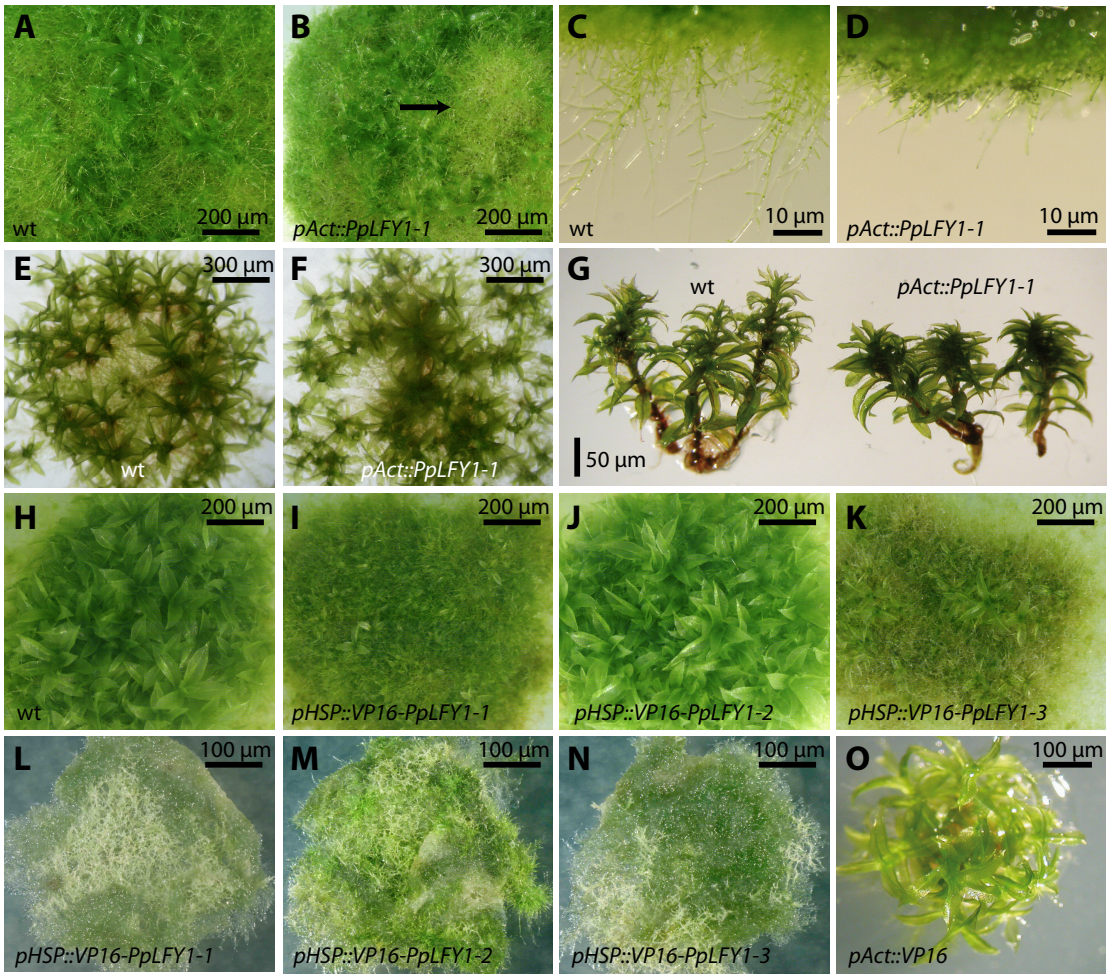


Figure 22: Phénotypes des clones de *P. patens* surexprimant *PpLFY1* ou *VP16-PpLFY1*. A-G: Phénotype du clone *pAct::PpLFY1-1* (**B, D, F, G droite**) en comparaison avec une plante sauvage (**A, C, E, G gauche**). **A-D:** Plantes ayant poussé sur milieu PpNH4 pendant 24 jours. Une zone de sénescence précoce des caulonémas et protonémas est indiquée par une flèche noire. Les caulonémas sont également beaucoup plus courts chez le surexpresseur. **E-F:** Plantes après un mois sur milieu PpNO3, *pAct::PpLFY1-1* développe des gamétophores. **G:** Gamétophores isolés après 67 jours de culture sur milieu PpNO3. Les gamétophores de *pAct::PpLFY1-1* (droite) apparaissent plus courts et trapus que ceux de la plante sauvage (gauche). **H-N:** Analyse du phénotype des 3 clones *pHSP::VP16-PpLFY1*. **H-K:** Clones *pHSP::VP16-PpLFY1-1* et *pHSP::VP16-PpLFY1-3* après 17 jours de culture sur milieu PpNH4, montrant une sénescence précoce des filaments. **L-N**: Phénotype des 3 clones après 15 jours de croissance sur milieu PpNH4, puis 3 jours de choc thermique (1h à 37°C par jour). Les clones *pHSP::VP16-PpLFY1-1* et *pHSP::VP16-PpLFY1-3* sont sévèrement affectés et la majorité des filaments entrent en sénescence. Le clone *pHSP::VP16-PpLFY1-2* montre un comportement similaire mais atténué. **O:** Plante *pAct::VP16* après 10 jours de culture.

Pour comprendre quels sont les gènes impliqués dans les phénotypes de surexpression de *PpLFY1*, nous avons envoyé les clones *pAct::PpLFY1-1* et *pAct::PpLFY1-2* pour une analyse par microarray à l'équipe de Stefan Rensing (Université de Freiburg) ; ces analyses sont actuellement en cours. Les plantes *pHSP::VP16-PpLFY1* seront analysées dans un deuxième temps si nécessaire.

Une fois obtenue la liste des gènes dérégulés chez le surexpresseur, nous pourrons rechercher lesquels de ces gènes possèdent un site de liaison pour PpLFY1 grâce à l'analyse bioinformatique précédente, et ainsi prédire un ensemble de gènes cibles directs de PpLFY1. La liaison de PpLFY1 à certains de ces gènes pourra éventuellement être validée par ChIP. Nous pourrons alors déterminer lesquels de ces gènes cibles sont communs avec ceux de LFY chez *A. thaliana*, ce qui pourrait représenter un réseau ancestral contrôlé par LFY. L'analyse des éléments *cis* au niveau de ces gènes communs nous montrera si les éléments *cis* peuvent « s'adapter » à un changement de spécificité, de telle sorte que la régulation du gène cible soit toujours préservée. Enfin, les gènes cibles spécifiques à PpLFY1 pourront nous éclairer sur le rôle particulier de LFY chez *P. patens*, et peut-être nous permettre de comprendre pourquoi PpLFY1/2 régule la première division cellulaire du zygote.

4) Evolution moléculaire du changement de spécificité LFY-PpLFY1

PpLFY1 et LFY reconnaissent des séquences très différentes *in vitro*. Quelle est la base moléculaire de cette différence de spécificité? Et comment ce changement de spécificité a-t-il pu avoir lieu pendant l'évolution, sans affecter ni les gènes cibles ni la fonction de LFY ?

a) Analyse structurale de la spécificité de liaison à l'ADN de PpLFY1

Deux membres de l'équipe, Camille Sayou et Renaud Dumas, en collaboration avec Max Nanao (EMBL, Grenoble), ont réussi à obtenir la structure cristallographique de PpLFY1-C (domaine C-terminal de PpLFY1) en complexe avec l'ADN, ce qui nous a permis d'identifier les différences avec la structure de LFY-C. Les deux protéines ont une organisation générale très similaire : PpLFY1-C lie l'ADN sous forme de dimère et le repliement d'un monomère se superpose très bien à celui d'un monomère de LFY-C (**Fig. 23A**). Par contre, PpLFY1-C établit un nouveau contact spécifique avec l'ADN, qui n'est pas présent chez LFY-C (**Fig. 23B**). En effet, à cet endroit LFY-C possède une histidine (H312)

qui est retenue éloignée de l'ADN par une arginine (R345), alors que PpLFY1-C possède respectivement aux mêmes positions un aspartate (D394) qui contacte une cytosine de l'ADN, et une cystéine (C427) qui n'influe pas sur la position de l'aspartate. L'aspartate D394 contacte directement la base C à la position -6 du motif d'ADN, et également légèrement la base C en position -7, les positions symétriques (+6 et +7) étant contactées par l'autre monomère (**Fig. 24C**). Ces positions du motif de liaison s'avèrent être les plus divergentes entre les motifs de LFY et de PpLFY1.

Les acides aminés D394 et C427 de PpLFY1 avaient déjà été identifiés par un alignement de séquences protéiques d'orthologues de LFY chez les plantes terrestres, puisqu'ils sont parfaitement conservés à part chez *P. patens* et *Atrichum angustatum* (une autre bryophyte proche de *P. patens*) (Maizel et al., 2005) (voir **Fig. 6** de l'introduction). La structure que nous avons obtenue explique leur importance pour le contact avec l'ADN et la spécificité de liaison de la protéine.

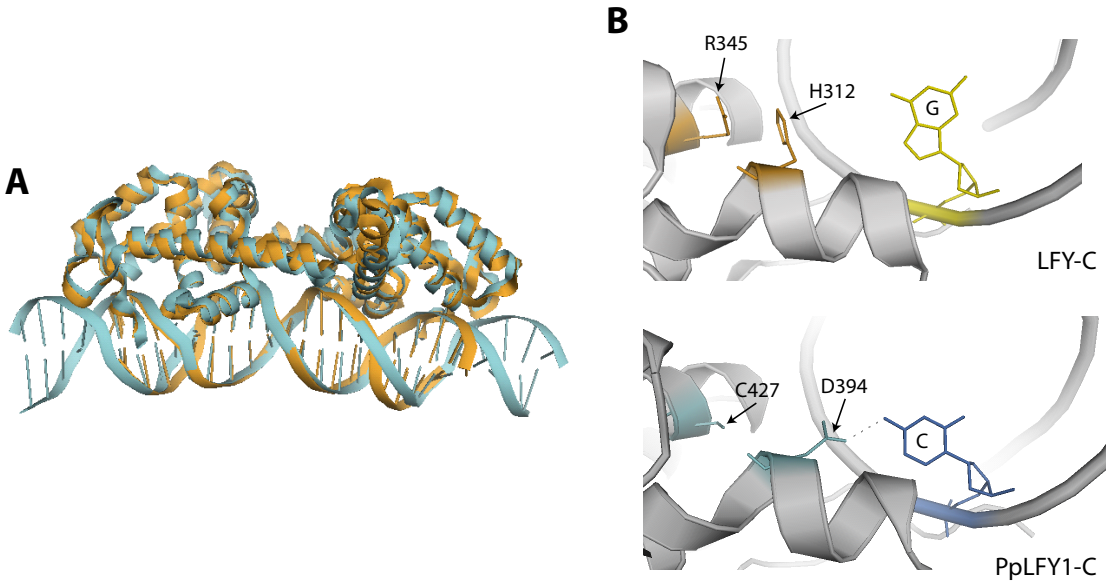


Figure 23 : Comparaison des structures de PpLFY1-C et LFY-C. A : Vue d'ensemble des deux structures superposées (orange : PpLFY1-C, bleu : LFY-C). **B :** Détail du contact à l'ADN différent entre les deux structures. Haut : LFY-C, la protéine ne contacte pas la guanine (G) de l'ADN à cette position, car l'histidine H312 est tenue éloignée de l'ADN par une arginine (R345). Bas : PpLFY1-C, la protéine contacte la cytosine (C) de l'ADN grâce à l'aspartate D394, qui n'est pas retenu par la cystéine C427.

La structure obtenue suggère que les deux acides aminés identifiés ont un rôle majeur dans la différence de spécificité des deux protéines, mais sont-ils suffisants pour expliquer cette différence ? Pour tester cela, Camille Sayou et moi-même avons produit des protéines recombinantes mutées au niveau d'un ou de ces deux acides aminés : nous avons ainsi créé une protéine de type LFYΔ avec les acides aminés de PpLFY1 aux deux positions à tester, et

vice-versa (**Fig. 24A**). Nous avons ensuite analysé ces protéines en gel retard pour déterminer leur spécificité de liaison (**Fig. 24B**).

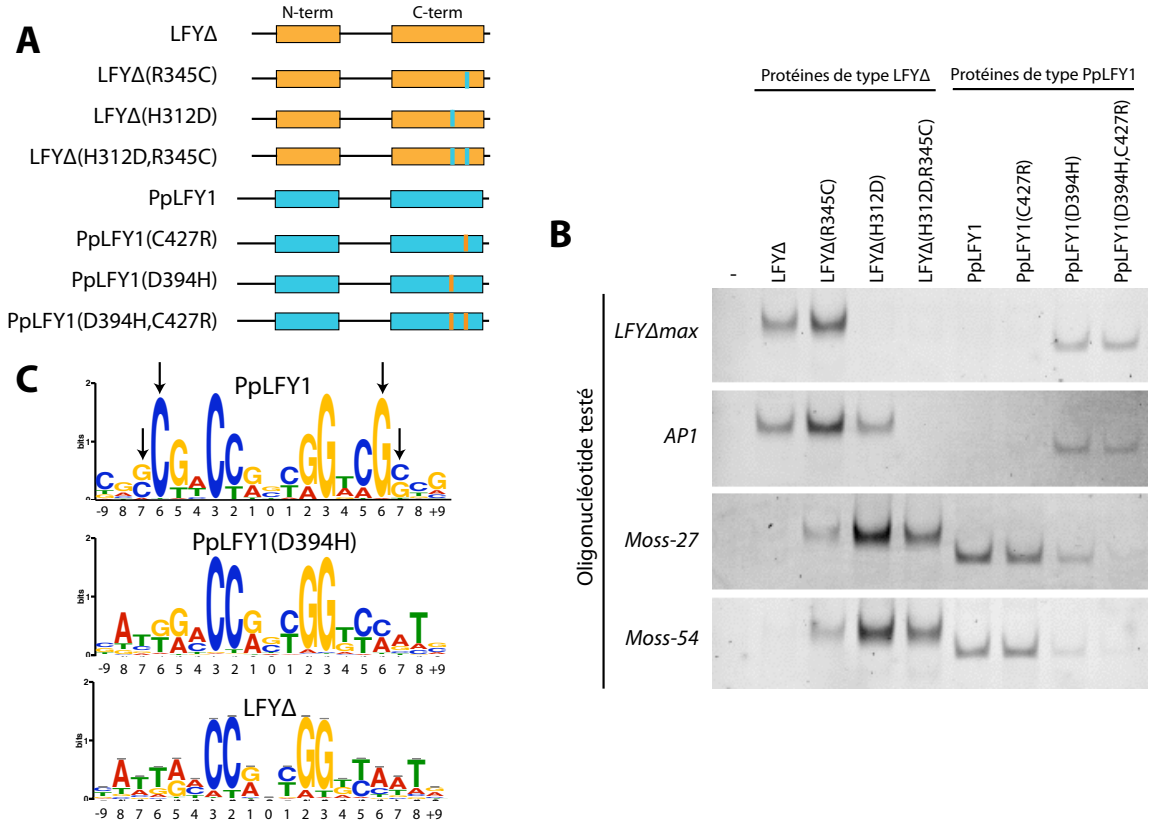


Figure 24 : Importance des acides aminés H/D et R/C pour la spécificité de liaison à l'ADN de LFYΔ et PpLFY1. A : Bilan des différentes protéines recombinantes produites. Les régions originaires de LFYΔ sont représentées en bleu, celles originaires de PpLFY1 sont en orange. **B :** Liaison des protéines purifiées sur 4 oligonucléotides représentant la spécificité de LFYΔ (*LFYΔmax* et *AP1*) ou de PpLFY1 (*Moss-27* et *Moss-57*). Seul le retard est représenté ici. Les protéines sont toutes à la même concentration (25 nM par piste). La première piste (-) ne contient pas de protéine. C : Logos obtenus après l'alignement des 2000 premières séquences uniques des SELEX de PpLFY1, PpLFY1(D394H) et LFYΔ. Le motif est contraint à 19 positions symétriques. Les positions contactées spécifiquement par l'aspartate D394 de PpLFY1 sont indiquées par des flèches noires.

Chez LFYΔ, la mutation R345C n'abolit pas la liaison de la protéine à l'oligonucléotide *LFYΔmax* (et semble même légèrement l'améliorer), mais permet la reconnaissance de l'oligonucléotide *Moss-27*, non reconnu par la protéine LFYΔ sauvage. Ce comportement est encore plus marqué pour la protéine LFYΔ(H312D) qui lie très fortement la séquence *Moss-27*, et ne reconnaît plus du tout *LFYΔmax*. La protéine LFYΔ(H312D,R345C) présente un comportement similaire. Les mutations de PpLFY1 confortent ces résultats: alors que PpLFY1 reconnaît *Moss-27* et non *LFYΔmax*, la protéine PpLFY1(D394H) lie les deux types de séquences, et la protéine PpLFY1(C427R,D394H) reconnaît uniquement *LFYΔmax*, adoptant alors une spécificité similaire à celle de LFYΔ. De plus, un SELEX avait été effectué

par Edwige Moyroud sur la protéine PpLFY1(D394H), et le motif de liaison obtenu est très proche de celui de LFYΔ (**Fig. 24C**), soulignant l'importance de cette mutation particulière sur la spécificité de la protéine entière.

Ainsi, en échangeant uniquement deux acides aminés (voire un seul selon les oligonucléotides considérés), nous avons pu complètement basculer la spécificité de LFY (que nous appellerons la spécificité HR) vers celle de PpLFY1 (spécificité DC), et vice-versa. Cette différence de spécificité n'est donc pas liée à d'importants remaniements structurels de la protéine, mais seulement à la séquence de deux acides aminés. De plus, Maizel et al. ont montré que, alors que PpLFY1 ne complémente pas un mutant *lfy* d'*A. thaliana*, la protéine PpLFY1(C427R,D394H) le complémente partiellement. Ceci signifie que, *in vivo*, la mutation de ces deux acides aminés chez PpLFY1 est suffisante pour rétablir l'expression de nombreux gènes chez *A. thaliana*, et donc que les éléments *cis* sont directement sensibles à ce changement de spécificité.

b) Un changement de spécificité soudain ?

Nous avons vu précédemment que LFY n'avait pas formé de famille multigénique au cours de l'évolution, et qu'il ne s'était que très rarement dupliqué. *P. patens* est un exemple d'espèce comprenant deux copies de *LFY* : *PpLFY1* et *PpLFY2*. Les deux protéines possèdent les acides aminés D et C aux positions identifiées précédemment, et sont très proches en séquence, ce qui suggère très fortement qu'elles ont la même spécificité de liaison à l'ADN. La présence de deux copies de *LFY* a-t-elle permis le changement de spécificité drastique que nous avons observé ? En effet, on peut imaginer que la présence de deux copies de *LFY* permettrait à l'une des copies de changer de spécificité alors que l'autre garde la spécificité ancestrale, assurant ainsi le contrôle des gènes cibles. Il existerait donc un état de transition avec deux protéines aux spécificités différentes, ce qui pourrait par exemple autoriser les éléments *cis* à s'adapter progressivement à la « nouvelle » spécificité, avant que la deuxième copie ne change aussi de spécificité. A l'opposé, si le facteur de transcription ne s'est pas dupliqué ou s'il s'est dupliqué après le changement de spécificité, celui-ci est brutal et la régulation des gènes cibles devrait alors être fortement perturbée de façon très soudaine.

Pour déterminer si la duplication de LFY chez *P. patens* est récente, nous avons cherché à replacer PpLFY1 et PpLFY2 sur un arbre phylogénétique, par rapport à deux copies de LFY (AtranFLO1 et AtranFLO2) chez *Atrichum angustatum*, une bryophyte proche. Ces deux protéines possèdent également les acides aminés D et C, et donc nous supposons

qu'elles ont la même spécificité que PpLFY1/2. Ces dernières sont très proches en séquence (seulement 17 acides aminés différents entre les deux protéines, et 95% d'identité) ; les protéines PpLFY1 et PpLFY2 sont donc toujours regroupées sur les arbres phylogénétiques, et associées à un noeud à forte valeur d'aLRT (approximate Likelihood Ratio Test, test proche de celui du bootstrap) (**Fig. 25**). Les séquences d'AtranFLO1 et AtranFLO2 sont légèrement plus divergentes entre elles, et ne seront pas toujours regroupées ensemble selon les contraintes appliquées à la phylogénie (**Fig. 25B**). La très forte ressemblance entre PpLFY1 et PpLFY2 suggère néanmoins fortement que la duplication de *LFY* chez *P. patens* est récente, et probablement postérieure au changement de spécificité, qui aurait eu lieu par conséquent alors qu'il n'y avait qu'une seule copie de *LFY*.

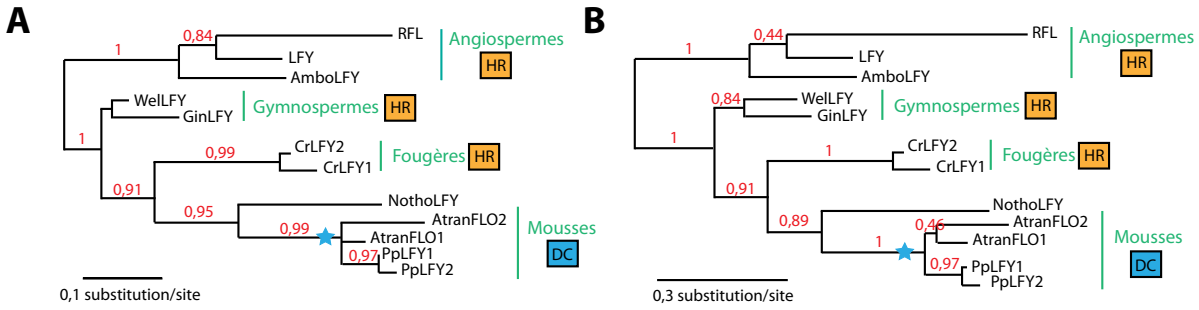


Figure 25 : Arbres phylogénétiques construits à partir de séquences entières de protéines LFY. Ces deux arbres montrent que PpLFY1 et PpLFY2 sont toujours regroupés, alors que la position d'AtranFLO1 et AtranFLO2 est plus variable. Les arbres sont obtenus à partir de l'alignement de 12 séquences d'orthologues de LFY issus des grands groupes de plantes terrestres. L'alignement est réalisé avec Muscle, est raffiné avec GBlocks (A) ou Built-in curer (B), puis l'arbre est construit avec PhyML (Maximum Likelihood). Les valeurs d'aLRT sont indiquées sur les branches. La séquence des deux acides aminés impliqués dans la spécificité de reconnaissance à l'ADN est indiquée à côté de chaque groupe (HR ou DC). L'étoile bleue positionne l'évènement de changement de spécificité supposé, conduisant à la spécificité actuelle de PpLFY1, PpLFY2, AtranFLO1 et AtranFLO2.

c) Vers la spécificité ancestrale de LFY

Nous ne savons pas pour l'instant « dans quel sens » s'est fait le changement de spécificité : la spécificité ancestrale était-elle plutôt du type HR, du type DC, ou était-ce une spécificité encore différente ? Pour déterminer cela, nous avons recherché d'autres séquences de *LFY* disponibles, et avons eu la chance de bénéficier de données transcriptomiques récentes dans la lignée verte, en collaboration avec Edwige Moyroud et Samuel Brockington (Université de Cambridge) ; ces données ne sont pas publiques pour l'instant et ont donc un caractère confidentiel. Les transcriptomes nous ont révélé que *LFY* n'apparaissait pas chez les plantes terrestres, comme supposé jusqu'à présent, mais qu'il était en réalité déjà présent chez un groupe plus basal de la lignée verte : les algues vertes. Nous avons ainsi pu identifier *LFY* depuis l'espèce *Klebsormidium* jusqu'aux plantes terrestres, mais pas chez deux algues vertes

unicellulaires dont les génomes sont séquencés, *Volvox* et *Chlamydomonas* (**Fig. 26A**). L'apparition de *LFY* paraît concorder avec l'apparition de la multicellularité, comme cela a été observé pour de nombreux facteurs de transcription (Rokas, 2008). Cette découverte totalement inattendue remet donc en question l'ordre des évènements ayant conduit aux spécificités HR et DC.

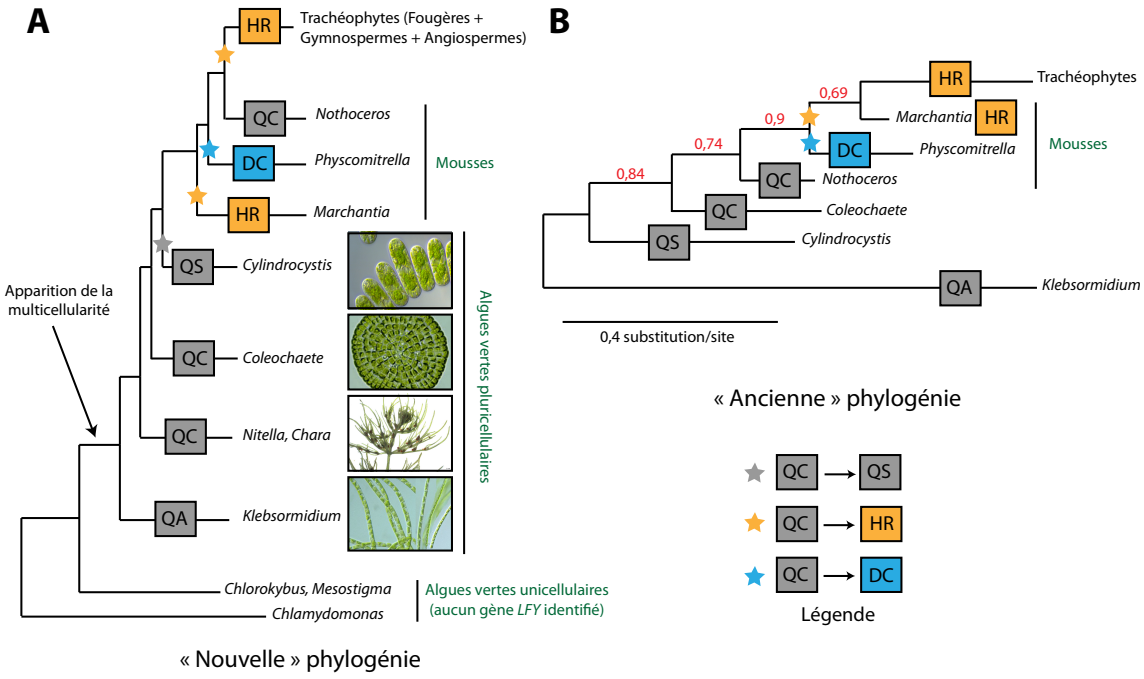


Figure 26 : Découverte de LFY chez les algues vertes et construction d'une phylogénie. A : « Nouvelle » phylogénie, inspirée de (Nishiyama and Kato, 1999) et (Timme et al., 2012). *Nothoceros* est l'espèce la plus proche des Trachéophytes, alors que *Marchantia* est à la base des plantes terrestres. L'algue verte *Cylindrocystis* est dans le groupe frère des plantes terrestres. Les acides aminés impliqués dans la spécificité de reconnaissance à l'ADN sont indiqués sur chaque branche. Depuis la protéine ancestrale LFY de type QC, il faudrait quatre évènements de mutations (symbolisés par des étoiles) pour obtenir les spécificités actuelles de LFY chez les mousses et les trachéophytes. **B :** Phylogénie construite à partir de 7 séquences entières de LFY (sauf pour MarpoFLO où le domaine C-terminal a été utilisé), l'alignement est effectué par Muscle, est raffiné par GBlocks, puis la phylogénie est calculée par PhyML avec les valeurs d'aLRT indiquées sur les branches en rouge. *Coleochaete* apparaît l'espèce d'algue verte la plus proche des plantes terrestres. *Nothoceros* est à la base des plantes terrestres, alors que *Marchantia* est la plus proche des Trachéophytes. Cette phylogénie ne nécessite que deux évènements de mutations pour obtenir les combinaisons actuelles observées chez les orthologues de LFY des mousses et des Trachéophytes.

Nous avons recherché les acides aminés impliqués dans la spécificité de reconnaissance à l'ADN, et nous avons identifié de nouvelles combinaisons de type QC (*Nothoceros*, *Coleochaete*), QS (*Cylindrocystis*) ou QA (*Klebsormidium*) (**Fig. 26A**). D'après les données structurales, nous imaginons que ces combinaisons seront équivalentes en termes

de spécificité de liaison, laquelle sera déterminée majoritairement par la glutamine ; nous noterons donc cette spécificité ancestrale QC.

Si l'on replace les différents types de spécificités en correspondance avec un arbre phylogénétique actuellement bien admis pour la lignée verte (Qiu, 2008a; Chang and Graham, 2011; Timme et al., 2012), on constate que le nombre d'évènements de mutations nécessaires pour expliquer les spécificités actuelles de LFY est de 4, ce qui est assez élevé, et comprend en outre deux évènements successifs de mutations QC > HR (**Fig. 26A**). Nous avons donc tenté de construire une phylogénie en utilisant les séquences protéiques de LFY dont nous disposions, ce qui nous a amené à proposer une phylogénie différente, réduisant le nombre d'évènements de mutations à 2 pour expliquer les spécificités actuelles (**Fig. 26B**). La phylogénie que nous avons établie était en fait admise il y a quelques années (ce pourquoi nous l'appelons « l'ancienne phylogénie »), et place le groupe des marchantiophytes comme le plus proche des trachéophytes (Malek et al., 1996; Nishiyama and Kato, 1999). Nous pensons que cette phylogénie est robuste car les séquences protéiques de LFY se prêtent particulièrement bien aux analyses phylogénétiques de par leur forte conservation de séquence, et car cet arbre est très parcimonieux pour expliquer les changements de spécificité observés pour LFY. La découverte de nouvelles séquences de LFY nous amène donc à remettre en question la phylogénie actuellement établie.

D'après les deux phylogénies, le changement de spécificité HR - DC n'a dans tous les cas jamais eu lieu, mais à sa place ont eu lieu deux changements : QC > DC et QC > HR. Il nous faut donc déterminer quelle est la spécificité de type QC pour comprendre l'ampleur de tels changements de spécificité. Pour cela, des expériences de SELEX sont en cours sur les protéines recombinantes NothoLFY (*Nothoceros*) et KlebsoLFY (*Klebsormidium)*. Nous imaginons que la spécificité QC est peut-être intermédiaire entre celle de LFY et celle de PpLFY, et qu'elle autorisera plus facilement une transition vers ces deux spécificités, de manière plus graduelle qu'un changement HR – DC direct, ce qui pourrait tout de même assurer le contrôle des gènes cibles ancestraux.

Nous avons étudié le changement de spécificité observé chez PpLFY1 d'un point de vue moléculaire, fonctionnel et évolutif. Ce changement est dû à la mutation de deux acides aminés uniquement, qui sont totalement responsables de la spécificité de la protéine in vitro*. L'impact de cette modification sur la régulation des gènes cibles de PpLFY1 est en cours d'analyse. Déterminer la spécificité de la protéine LFY ancestrale chez les algues vertes nous permettra de comprendre l'historique de ce changement de spécificité, et s'il a été brutal ou non. L'étude de LFY chez* Physcomitrella patens *se révèle donc très riche pour comprendre les mécanismes d'évolution d'un facteur de transcription et de ses cibles.*

DISCUSSION ET PERSPECTIVES

I) LFY et son évolution dans la lignée verte : une histoire originale et variée

Les possibilités d'évolution du réseau contrôlé par LFY sont très variées. Si l'on se focalise sur la partie du réseau reliant LFY à ses gènes cibles, on peut imaginer des variations au niveau des éléments *cis* des gènes cibles ou au niveau de la séquence codante de LFY lui-même. Le nombre de copies de *LFY* étant toujours resté faible et la séquence de son domaine de liaison à l'ADN étant extrêmement conservée, les possibilités d'évolution en *trans* paraissaient a priori très faibles. En étudiant la protéine LFY chez différentes espèces, nous avons pourtant identifié plusieurs cas de modifications en *trans* du réseau (PpLFY et LFY, LFY et NLY chez les gymnospermes) ainsi que de nombreux cas d'évolution en *cis* (profils de scores des introns d'*AG* ou *SHP* chez les angiospermes). Tout ceci participe vraisemblablement à la grande variabilité des situations observées quant au rôle de LFY *in planta*, en comparaison à celui décrit chez *A. thaliana* : parfois ce rôle est parfaitement conservé (chez *Antirrhinum* par exemple), parfois seulement partiellement (chez le riz ou le tabac, les orthologues de *LFY* contrôlent également l'architecture de l'inflorescence), ou parfois le rôle de LFY apparaît totalement divergent (chez *P. patens*). Même au sein des angiospermes où *LFY* a pourtant un rôle essentiel et globalement conservé pour le développement floral, il existe toujours de légères variations de ce rôle ou des manières dont il s'accomplit chez chacune des espèces que l'on étudie. Je vais donc discuter de ces multiples chemins choisis par l'évolution pour faire varier un réseau génétique.

1) Un gène unique

a) *LFY* sans famille

Nous avons vu précédemment que *LFY* était présent majoritairement en une ou deux copies (rarement 3, comme chez le pommier) dans l'ensemble des génomes des plantes terrestres (**Fig. 5** de l'introduction). Ainsi, *LFY* n'a jamais été retenu en plusieurs copies après les évènements de duplication du gène lui-même qui ont pu se produire, ni après les nombreux évènements de duplications entières du génome qui ont eu lieu au cours de

l'évolution (3 chez *Arabidopsis* depuis l'émergence des angiospermes) (Adams and Wendel, 2005). A l'exception, certains des cas où *LFY* a été retenu à l'état de 2 copies correspondent à des évènements récents de polyploïdisation comme chez le maïs ou le tabac (dernier événement de duplication daté respectivement de 11-14 millions d'années et 200 000 ans) (Adams and Wendel, 2005; Leitch et al., 2008).

Pourquoi *LFY* n'est-il jamais retenu en de nombreuses copies ? L'une des hypothèses est que la présence de plusieurs copies de *LFY* soit délétère au développement de la plante et qu'elles soient donc éliminées par sélection naturelle. En effet, introduire plusieurs copies de *LFY* chez *A. thaliana* résulte en une floraison légèrement précoce, ce qui est également observé, beaucoup plus fortement, chez un surexpresseur *35S::LFY* (Weigel and Nilsson, 1995; Blazquez et al., 1997). Or, une floraison précoce peut se révéler létale pour une plante dans la nature, puisque l'attente des conditions environnementales propices à une floraison réussie est cruciale. Une autre hypothèse, proposée par Baum et al., est que les copies de *LFY* ne se subfonctionnalisent ou ne se néofonctionnalisent pas, et sont donc perdues par dérive génétique (Baum et al., 2005). C'est en effet ce que nous observons pour les copies de *LFY*, qui ne montrent jamais de subfonctionnalisation véritable (à part chez les gymnospermes), mais tout au mieux une redondance partielle. Chez l'eucalyptus, une seule des deux copies de l'orthologue de *LFY* (*ELF1*) est fonctionnelle, la deuxième étant un pseudogène (Southerton et al., 1998). Chez le maïs, les deux gènes *ZFL1* et *ZFL2* sont partiellement redondants, la mutation *zfl1* étant spécifiquement associée à une simplification des ramifications de l'inflorescence (Bomblies et al., 2003). Enfin, chez *P. patens*, il est possible que la duplication de *PpLFY* soit reliée à un événement de polyploïdisation assez récent (30-60 Ma) (Rensing et al., 2007), ce qui pourrait expliquer pourquoi les deux copies de *PpLFY* sont encore maintenues. *PpLFY1* et *PpLFY2* sont également partiellement redondants : le simple mutant *pplfy1* présente un phénotype plus marqué que celui de *pplfy2* (Tanahashi et al., 2005) mais moins que celui du double mutant *pplfy1 pplfy2* . On peut imaginer, mais cela n'est bien sûr qu'une hypothèse, que *PpLFY2* va tendre à disparaître puisque *PpLFY1* semble remplir l'essentiel de la fonction biologique des deux gènes.

Le seul cas de duplication de *LFY* qui paraisse associé à une véritable subfonctionnalisation des gènes, et qui ait été retenu chez de nombreux taxons et pendant une longue période évolutive, est celui de *LFY* et *NLY* chez les gymnospermes. Comme nous l'avons vu dans l'Article 4, le contrôle des gènes B et C semble être partagé entre LFY et

NLY chez *W. mirabilis*, et d'autres indices sur le domaine d'expression de ces gènes chez d'autres gymnospermes suggèrent également leur subfonctionnalisation. Ceci peut expliquer pourquoi *LFY* et *NLY* ont été retenus chez les gymnospermes, sans néanmoins justifier pourquoi *NLY* a été perdu chez les angiospermes.

b) Evoluer sans se dupliquer

Le nombre de copies de *LFY* est peu variable, et les facteurs de transcription évoluent fréquemment par duplication ; cela signifie-t-il que LFY ne peut évoluer en *trans* qu'à l'état de deux copies ? L'arbre phylogénétique construit en Fig. 26B, qui intègre les séquences de LFY chez les algues vertes, nous montre qu'il y aurait eu deux changements de spécificité majeurs au cours de l'évolution de LFY : un changement QC > DC sur la lignée des bryophytes, et un changement QC > HR sur la lignée des marchantiophytes et des trachéophytes. Nous n'avons trouvé qu'une seule copie de *LFY* dans le transcriptome des algues vertes, suggérant qu'il n'existe qu'un seul gène *LFY* chez ces espèces. Les deux changements de spécificité proposés se seraient donc tous les deux produits alors qu'une seule copie de *LFY* était présente.

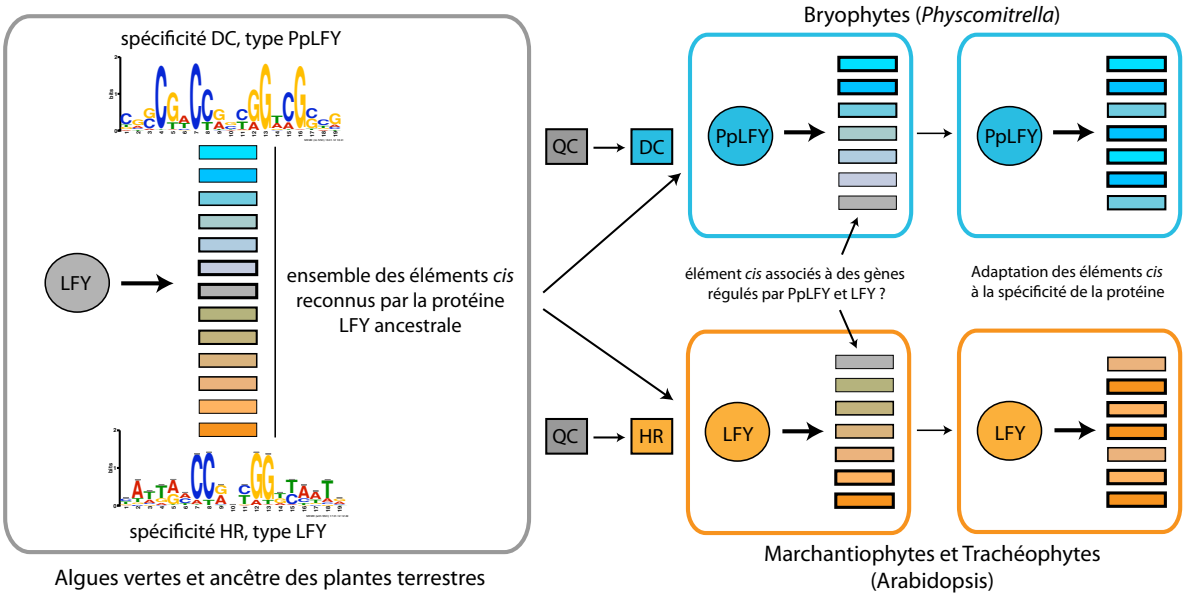


Figure 27 : Modèle des changements de spécificité de LFY et de leur impact modéré sur la liaison de LFY à ses éléments *cis*. On suppose que la protéine LFY ancestrale (en gris) reconnaissait un ensemble d'éléments *cis* de spécificités intermédiaires entre HR et DC. Les éléments *cis* sont symbolisés par des rectangles de couleur, leur contour étant plus épais si leur affinité pour LFY est grande. Lors des changements de spécificité QC > DC et QC > HR, seule une sous-partie de ces éléments *cis* ont continué à être reconnus par les protéines dans chacun des cas, maintenant ainsi la régulation d'une partie des gènes cibles. On peut imaginer que les éléments *cis* encore reconnus par PpLFY et LFY actuellement sont associés à des gènes régulés par les deux protéines, et reflètent peut-être une fonction ancestrale de LFY. Eventuellement, les éléments *cis* se seront ensuite adaptés à la spécificité de la protéine.

De tels changements de spécificité pourraient avoir lieu si leurs conséquences ne sont pas dramatiques, ce qui est envisageable si l'on suppose que la spécificité ancestrale de LFY était intermédiaire entre celle de PpLFY et celle de LFY. Dans ce cas, on peut imaginer que la protéine LFY ancestrale reconnaissait un ensemble d'éléments *cis* correspondant à la spécificité HR, à la spécificité DC, et à tous leurs intermédiaires (**Fig. 27**). Après les mutations QC > DC, la protéine ne reconnaissait plus qu'une sous partie des éléments *cis*, assurant ainsi la régulation de certains gènes seulement ; un processus similaire se serait déroulé lors des mutations QC > HR. Ainsi, les changements de spécificité n'auraient pas eu comme conséquence la perte totale et soudaine de la régulation de tous les gènes par LFY, mais seulement d'une sous-partie de ces gènes. On peut imaginer que les gènes possédant des éléments *cis* de spécificité intermédiaire entre DC et HR chez l'ancêtre commun correspondraient aux gènes actuellement encore régulés de façon commune par LFY et PpLFY, par exemple des gènes contrôlant la division cellulaire ou un état méristématique (voir Introduction, III-2). Il est possible que les éléments *cis* se soient ensuite adaptés aux mutations des protéines LFY pour mieux correspondre à leur spécificité de liaison.

Pour vérifier l'hypothèse selon laquelle la protéine LFY ancestrale avait une spécificité intermédiaire entre celle de PpLFY et celle de LFY, des expériences de SELEX sont en cours sur les protéines LFY de *Nothoceros* et de *Klebsormidium*. Des expériences de complémentation du mutant *lfy* d'*A. thaliana* et du double mutant *pplfy1 pplfy2* de *P. patens* pourraient être envisagées pour déterminer l'impact de cette spécificité intermédiaire, si elle existe, *in vivo*.

2) Evolution *cis* ou *trans* ?

a) Des motifs similaires en SELEX

Nous avons employé la technique de SELEX pour déterminer si LFY avait évolué en *trans*, car c'est une technique relativement simple et rapide, applicable à n'importe quel facteur de transcription. Comment savoir si le motif obtenu, qui est toujours le même pour presque toutes les protéines LFY, ne correspond pas à un artéfact expérimental ou à un biais de l'alignement des séquences ? En effet, le SELEX présente quelques défauts, inhérents au principe même de la technique : des bordures constantes sont présentes dans la librairie et pourraient être liées par la protéine ; l'amplification des oligonucléotides par PCR à chaque cycle peut être biaisée,... De plus, le choix du cycle auquel arrêter l'enrichissement est

crucial : une sélection trop faible engendrera un bruit de fond de liaison aspécifique important, alors qu'une sélection trop poussée risque de n'isoler que quelques séquences de très bonne affinité pour LFY, et donc de perdre la diversité attendue pour sa spécificité de liaison. Pourtant, nous pensons que les résultats du SELEX sont robustes pour plusieurs raisons : (1) la séquence au meilleur score prédit, ainsi que des séquences ne montrant qu'une mutation par rapport à celle-ci, sont retrouvées de très nombreuses fois dans le jeu de séquences issues du SELEX, ce qui témoigne d'un enrichissement efficace ; (2) pour LFYΔ, un arrêt de l'enrichissement à 2, 3 ou 4 cycles de SELEX génère des motifs de liaison très similaires (**Fig . 12**), soulignant la robustesse de l'algorithme de MEME ; (3) le motif obtenu pour LFYΔ a pu être confirmé *in vitro* par gel retard ; (3) ce même motif a été obtenu par des techniques *in vivo* de ChIP-Seq (**Article 2**) ou de ChIP-chip (**Article 3**).

Les logos que nous avons obtenus en SELEX sont très proches pour toutes les protéines LFY, excepté celle de *P. patens*. Pourtant les fréquences des nucléotides à chaque position ne sont pas strictement équivalentes ; comment savoir si ces légères différences auront un impact significatif sur la reconnaissance de la protéine à l'ADN ? La famille des facteurs de transcription ETS-1, aux fonctions physiologiques variées chez l'homme, a été extensivement étudiée pour la spécificité de liaison de ses différents membres. Une étude a déterminé la spécificité de 27 de ces facteurs par Protein Binding Microarray (voir Introduction), laquelle s'est révélée très conservée, notamment le coeur central du motif (Wei et al., 2010). Pourtant, les auteurs ont pu identifier, grâce à un programme d'alignement des matrices qu'ils ont développé, des différences significatives dans ces motifs de liaison ; ces différences sont localisées majoritairement à une seule des positions du site de liaison de 10 pb, et conditionnent les séquences reconnues par ces protéines en ChIP-Seq. Le programme STAMP, similaire à celui développé par Wei et al., permet de comparer des matrices de liaison, en les alignant pour calculer une phylogénie à partir de cet alignement (Mahony and Benos, 2007). J'ai utilisé ce programme pour tenter de détecter des différences entre les matrices issues du SELEX (**Fig. 28**). Alors que le programme distingue bien que la matrice de PpLFY1 et celles des protéines NLY sont différentes du reste des matrices, aucune distinction n'est faite entre les matrices des autres protéines LFY. Ceci nous conforte donc dans l'idée que les spécificités des protéines LFY, à l'exception de celle de PpLFY1, ne montrent pas de différences significatives entre elles.

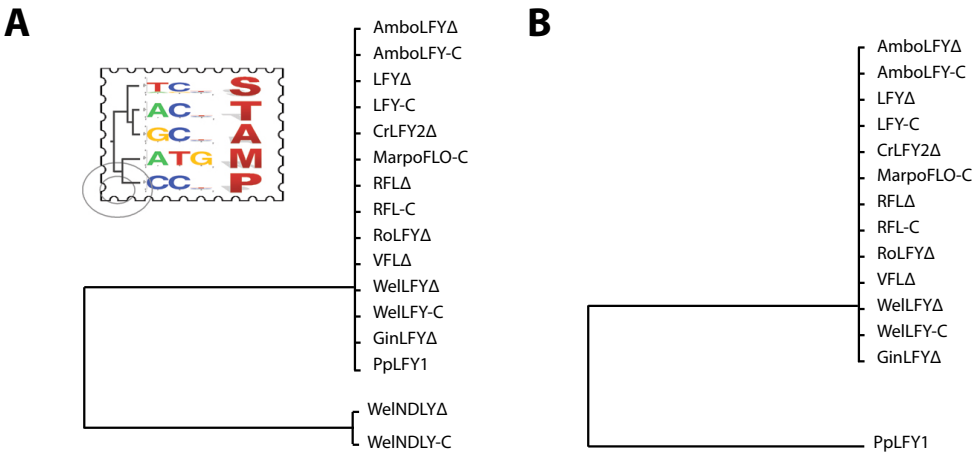


Figure 28 : Résultat des comparaisons des matrices issues du SELEX par le programme STAMP. A : Les matrices de fréquences de toutes les protéines étudiéesont été comparées, seules les matrices de WelNDLYΔ et WelNDLY-C sont différenciées des autres matrices. **B :** Même analyse, en excluant les matrices de WelNDLYΔ et WelNDLY-C : le programme détecte des différences entre la matrice de PpLFY1 et le reste des matrices. Si la matrice de PpLFY1 est exclue de cette analyse, aucune différence n'est faite entre le reste des matrices.

b) Des changements en *trans*, des changements en *cis*

Nous avons vu que le réseau contrôlé par LFY avait évolué en *trans* ou en *cis* selon les espèces considérées. Pourtant, le fait de détecter des changements *cis* ou *trans* ne signifie pas obligatoirement que la fonction du facteur de transcription et les gènes qu'il régule aient varié. Chez les levures *Saccharomyces cerevisiae* et *Candida albicans*, la détermination des cellules en type a/α pour la reproduction est déterminée respectivement par une régulation négative ou positive des gènes spécifiques du type a (gènes asgs). Cette différence de régulation est associée à des mutations au niveau des éléments *cis* des gènes asgs ainsi qu'à des changements en *trans* chez les facteurs de transcription MATa2 et MATα2 régulant ces gènes ; les deux situations conduisent pourtant à une régulation identique des gènes asgs et à une différenciation similaire des types cellulaires a et α (Tsong et al., 2006). Une autre étude plus récente chez la levure *Candida glabrata* montre la coévolution possible entre éléments *cis* et *trans* dans le cas du réseau contrôlé par le facteur CgAP1 (Kuo et al., 2010). Ce facteur et ses orthologues chez des levures proches reconnaissent des éléments de liaison différents *in vitro*, ce qui dépend d'une mutation en *trans* du facteur, mais ils lient pourtant un ensemble de gènes communs *in vivo*, participant ainsi à la réponse aux stress. Les auteurs ont pu identifier des mutations dans les éléments *cis* associés à ces gènes chez *C. glabrata*, qui sont alors reconnus par CpAG1. Kuo et al. proposent que ce mécanisme de coévolution entre mutations *cis* et *trans* est vraisemblablement favorisé par la présence de paralogues proches de CpAG1 qui assurent la fonction de la protéine alors qu'elle a changé de spécificité de liaison, laissant

ainsi le temps aux éléments *cis* de s'adapter en retour. Des effets compensatoires à grande échelle sont donc possibles entre des mutations *cis* et *trans*. Il est difficile d'imaginer ce genre d'évènement dans le cas de *LFY*, gène à la fonction essentielle, et en copie unique pendant les changements de spécificité que nous avons identifiés. Il est tout de même possible qu'une partie des gènes soient toujours régulés de manière commune par LFY et PpLFY, si les éléments *cis* se sont adaptés au changement de spécificité de la protéine.

A l'inverse, le fait de ne pas détecter de changement en *trans* entre les différentes protéines LFY des angiospermes ne signifie pas nécessairement que les gènes régulés par ces protéines seront les mêmes. Comment expliquer que RFL, l'orthologue de LFY chez le riz, contrôle les ramifications de l'inflorescence et l'émergence des talles (Rao et al., 2008), alors que sa spécificité est la même que celle de LFY, qui ne contrôle pas ces aspects développementaux ? Comme proposé dans l'Article 1, il est vraisemblable que la simple modification du domaine d'expression de *RFL* par rapport à celui de *LFY* lui ait permis d'acquérir de nouvelles fonctions. En effet, *RFL* est exprimé de façon transitoire au niveau des méristèmes de branchement de l'inflorescence et dans les régions d'émergence des talles (Kyozuka et al., 1998; Rao et al., 2008), ce qui lui permet potentiellement d'interagir avec de nouveaux corégulateurs (Ikeda et al., 2007), ou encore de se lier à des éléments *cis* accessibles spécifiquement dans ce tissu, et donc de contrôler l'expression de gènes différents. La modification de la spécificité de liaison d'un facteur de transcription n'est donc pas le seul moyen pour qu'il reconnaisse des éléments *cis* différents.

Le changement en *trans* que nous avons observé entre LFY et PpLFY1 ne repose, au niveau moléculaire, que sur la mutation de deux acides aminés. Ainsi, deux mutations seulement peuvent engager un facteur de transcription dans la voie de contrôle d'un ensemble de gènes, et dans l'acquisition d'une fonction spécifique. Par opposition, les mutations au niveau des éléments *cis* sont plus nombreuses, soulignant la fluidité en *cis* du réseau, déjà observée à plusieurs reprises (Borneman et al., 2007; Schmidt et al., 2010), et ces mutations semblent avoir des conséquences beaucoup plus discrètes. Le profil de liaison de LFY aux introns d'*AG* chez les angiospermes est très variable, pourtant de nombreux indices suggèrent que LFY régule tout de même *AG* chez les espèces étudiées. De plus, comme nous l'avons vu en comparant le cas d'*Arabidopsis*, d'*Antirrhinum* et des Rosacées, le « choix » du facteur qui accomplira la fonction C et sera régulé par LFY est variable ; pourtant dans tous les cas une fleur fertile, organisée de façon similaire, se développera. Ainsi, alors que les changements en *trans* que nous avons observés semblent engager LFY dans une voie de régulation restreinte

et spécifique, les changements en *cis* chez les angiospermes paraissent plutôt nuancer le mode d'action de LFY, qui conserve alors l'essentiel de sa fonction. Il n'existe au final jamais deux réseaux contrôlés par LFY qui soient totalement identiques, même au sein des angiospermes.

II) Prédire des gènes cibles et une fonction ancestrale pour LFY ?

Comprendre pourquoi un ensemble de gènes est exprimé spécifiquement dans un tissu est un enjeu majeur en biologie, et de très nombreuses études s'attachent à détecter la présence d'éléments régulateurs en amont des gènes pour répondre à cette question. Pour cela, des modèles prédictifs de la liaison de facteurs de transcription sur une séquence génomique sont développés, qui peuvent ensuite être appliqués aux cas d'espèces non modèles pour comprendre l'évolution des régulations transcriptionnelles, et inférer l'apparition d'une régulation ou d'une fonction au cours de l'évolution. Je vais discuter du modèle que nous avons développé pour prédire la liaison de LFY-C à l'ADN, en relation avec les avancées récentes dans ce domaine, et je vais aborder comment ce modèle pourrait nous aider à découvrir la fonction ancestrale de LFY.

1) Prédire une liaison, prédire une régulation

a) Un modèle hautement prédictif

Le modèle biophysique que nous avons développé nous permet de prédire les sites de liaison de LFY-C sur une séquence génomique. Pour évaluer la performance de ce modèle, nous avons construit une courbe de type ROC (Receiver Operating Characteristic) qui compare le taux de faux positifs (sites prédits par le modèle, mais non liés en ChIP-Seq) à celui de vrais positifs, selon un seuil de POcc variable (**Article 2, Figure 3**) (Clarke and Granek, 2003; Granek and Clarke, 2005). Un modèle parfaitement prédictif aura un taux de vrais positifs de 1, et un taux de faux négatifs de 0 ; ceci engendrera une courbe en marche d'escalier, et l'aire du domaine sous la courbe (ROC-AUC : ROC Area Under the Curve) sera de 1. Notre modèle offre une valeur de ROC-AUC de 0,865 ; par comparaison, seulement 11 des 28 facteurs de transcription analysés par Roider et al. présentent une valeur de ROC-AUC égale ou supérieure, en utilisant pourtant un calcul similaire d'Occupation Prédite (Roider et al., 2007). Nous pensons que notre modèle offre un pouvoir prédictif particulièrement élevé

car la description de la spécificité de liaison de LFY-C est très précise, d'une part car de nombreuses séquences de SELEX sont utilisées pour l'alignement, et d'autre part car nous rajoutons une étape supplémentaire d'affinement de la matrice de liaison par la technique du QuMFRA, ce qui nous fournit un modèle reflétant très précisément la liaison de LFY-C *in vitro*. Pourtant, le motif de liaison de LFY-C n'offre pas un contenu en information particulièrement important (11,7 bits contre une moyenne de 11,9 bits pour les facteurs de transcription déposés sur la base de données JASPAR) (Sandelin et al., 2004). Par contre, il présente peu de positions « obligatoires » (c'est-à-dire avec une information proche de 2 bits) alors que de nombreuses positions montrent une information faible, qui ne dépasse pas 0,7 bits. Ainsi, pour prédire les sites de liaison de LFY-C, l'utilisation de la séquence consensus correspondant à ce motif serait particulièrement inadaptée ; au contraire l'utilisation d'une matrice poids/position décrit beaucoup plus précisément l'éventail de sites de liaison possibles pour LFY-C.

b) Un grand pas jusqu'à la régulation

Notre modèle de liaison pourrait encore être amélioré de nombreuses manières, pour intégrer des niveaux de complexité supplémentaires afin de prédire la liaison de LFY *in vivo*, et éventuellement les gènes qu'il régule (**Fig. 29**). Tout d'abord, il serait intéressant de détecter les sites de liaison de LFY sur l'ADN génomique d'*A. thaliana*, après avoir éliminé les histones et les facteurs chromatiniens. Une telle technique, baptisée PB-Seq (Protein/DNA Binding), a été appliquée au Heat Shock Factor (HSF) de la Drosophile (Guertin et al., 2012). Cette approche, qui utilise de longs fragments d'ADN à l'opposé du SELEX qui implique de petits oligonucléotides, permet de détecter des interactions à longue distance entre sites de liaison. Ce type d'interactions a déjà été mis en évidence pour plusieurs facteurs de transcription, comme pour les facteurs MADS qui peuvent faire un « looping » de l'ADN (formation d'une boucle) (Revet et al., 1999; Melzer and Theissen, 2009). Dans un deuxième temps, l'impact des facteurs chromatiniens, des histones, ou d'autres facteurs de transcription sur la liaison de LFY à l'ADN pourrait être intégré au modèle. Guertin et al., après avoir réalisé leur PB-Seq, ont croisé leurs données avec des résultats d'hypersensibilité à la DNase (qui reflète l'accessibilité à l'ADN), de positionnement des histones et de leurs modifications, ainsi que de la liaison d'autres facteurs de transcription. De façon intéressante, l'accumulation de toutes ces données engendre un modèle performant (corrélation entre les données de liaison en ChIP-Seq et les sites prédits : r=0,62), mais moins que la combinaison de deux de

ces facteurs seulement (r=0,7), l'hypersensibilité à la DNAse seule améliorant la corrélation de manière importante. Dans le cas du modèle de liaison de LFY, intégrer des données d'hypersensibilité à la DNase du génome *A. thaliana,* ainsi que le positionnement des nucléosomes, est tout à fait envisageable (Zhang et al., 2012). Ces données, ainsi que le positionnement d'autres facteurs de transcription comme AP1 ou SEP3, pourraient très vraisemblablement engendrer un modèle encore plus performant de prédiction de la liaison de LFY à ses sites, puisque de nombreux sites liés par LFY en ChIP-Seq le sont aussi par AP1 et/ou SEP3 (Kaufmann et al., 2009; Kaufmann et al., 2010b; Moyroud et al., 2011; Winter et al., 2011), suggérant une compétition ou une coopération entre ces différents facteurs pour les mêmes sites de liaison.

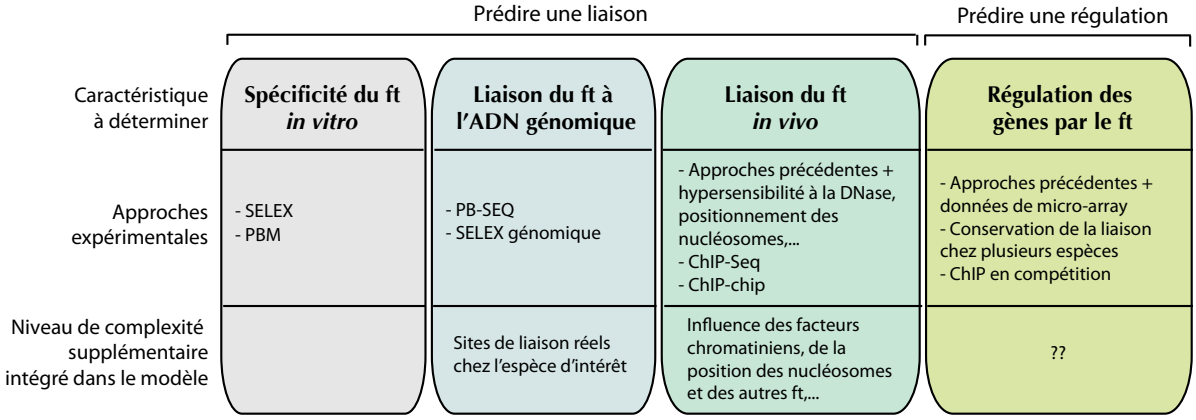


Figure 29 : Etapes pour aller de la prédiction de la liaison à l'ADN d'un facteur de transcription (ft) jusqu'à la prédiction de la régulation de ses gènes cibles. Les approches expérimentales développées pour chacune des étapes sont indiquées, ainsi que le niveau de complexité apporté par chaque étape par rapport à l'étape précédente. Les points d'interrogations au niveau de la prédiction de la régulation signifient que l'on ne sait pas pour l'instant de façon claire ce qui conditionne le passage d'une liaison *in vivo* à une régulation.

Les principes qui gouvernent le passage d'une liaison à une régulation sont par contre encore mal compris. J'ai tenté de prédire les gènes régulés par LFY en utilisant la conservation de sa liaison, prédite par la valeur de POcc, entre les génomes d'*A. thaliana* et d'*A. lyrata* ; ce type d'approches inter-espèces est prometteur et a été appliqué avec succès à grande échelle pour détecter la présence de modules régulateurs par la comparaison des génomes de 12 drosophiles (Stark et al., 2007). Il est probable que le nombre de génomes utilisés pour cette analyse soit déterminant pour augmenter le pouvoir prédictif du modèle, pourtant l'utilisation d'uniquement deux génomes de drosophiles proches a permis d'améliorer la prédiction des gènes impliqués dans la segmentation de l'embryon (Sinha et al., 2004). Pour prédire ces gènes, les auteurs n'ont pas utilisé la matrice poids/position d'un seul facteur de transcription, mais celles de tout un ensemble de facteurs impliqués dans le processus de segmentation, ce qui leur fournit un modèle avec environ 50 % de faux positifs,

donc plus performant que notre modèle qui en produit au minimum 75 % (Rajewsky et al., 2002; Sinha et al., 2004). Les gènes identifiés avec les deux génomes de drosophiles se sont révélés plus fortement exprimés dans les territoires reliés à la segmentation que les gènes identifiés avec un seul génome. Je n'ai pourtant pas obtenu de résultat similaire avec mes analyses chez *A. thaliana* et *A. lyrata*, puisque les niveaux de dérégulation des gènes chez *lfy-12* ou lors de la transition florale étaient comparables en utilisant un ou deux génomes. Il est donc difficile de comprendre pourquoi cette analyse n'a pas marché, alors que la distance évolutive entre les deux espèces d'*Arabidopsis* utilisées paraît adaptée à ce type d'étude. En analysant quelques gènes individuels, j'ai pu remarquer qu'*AP1*, qui montre une région à très forte valeur de POcc, était toujours retrouvé dans les gènes au-dessus du seuil imposé, alors que *TFL1* ou *AG* ne l'étaient pas systématiquement. En effet, les scores de liaison de LFY-C pour ces deux gènes ne sont pas très élevés (voir **Article 2, Figure 4**) et répartis sur une région assez grande. Il est donc possible que le calcul de la POcc sur 150 pb soit trop limitant pour de tels gènes et qu'il faudrait intégrer les scores sur une région plus étendue ; le temps de calcul informatique deviendrait par contre très important. La conservation de la liaison d'un ou plusieurs facteurs au cours de l'évolution reste dans tous les cas une piste importante pour prédire l'existence d'une régulation.

Une récente étude sur le facteur de transcription Rap1 chez *Saccharomyces cerevisiae* a montré de façon élégante que le lien entre liaison d'un facteur de transcription et régulation du gène pouvait dépendre de la dynamique de cette liaison (Lickwar et al., 2012). Les auteurs ont réalisé des expériences de ChIP-chip sur la protéine Rap1-Flag exprimée de manière constitutive, et mise en compétition avec une autre protéine Rap1-Myc exprimée de manière inductible à différents temps. Ainsi, ils ont pu mesurer la dynamique de la liaison de Rap1 à ses sites, selon la vitesse à laquelle le facteur était exclu du site par son compétiteur induit. Il en résulte que les sites où Rap1 reste présent le plus longtemps sont associés à des gènes fortement régulés par Rap1, alors qu'au niveau de gènes faiblement régulés par Rap1 le facteur montre un turn-over très rapide. La différence entre ces sites résiderait dans l'affinité de Rap1 pour le site, mais surtout dans leur occupation par les nucléosomes (faible pour des sites où Rap1 réside longtemps), laquelle serait déterminée par la séquence d'ADN elle-même. Les auteurs proposent alors que le changement de régulation d'un gène par un facteur de transcription, en fonction des conditions environnementales ou du tissu, soit contrôlé par un remaniement de la position des histones, vraisemblablement via des modifications post-traductionnelles. Le lien entre liaison et régulation est donc peut-être à chercher au niveau de

la dynamique de liaison du facteur de transcription à ses cibles. De manière générale, de nombreux progrès sont accomplis dans le domaine de la prédiction de la liaison et de la régulation d'un facteur de transcription à ses gènes cibles, et le « futility theorem », selon lequel la grande majorité des sites de liaison prédits pour un facteur de transcription n'ont en réalité aucun rôle biologique, se révèlera peut-être erroné au fur et à mesure que les modèles s'affineront (Wasserman and Sandelin, 2004).

2) La fonction ancestrale de LFY : des prédictions aux expérimentations

Notre modèle biophysique de liaison de LFY à l'ADN a pu être utilisé dans le cas de PpLFY1, pour prédire ses sites de liaison au génome de *P. patens*. Pour avoir une idée de la fonction des gènes associés à ces sites de liaison, j'ai recherché celle de leurs orthologues chez *A. thaliana*. De nombreux gènes impliqués dans la division cellulaire ou l'identité méristématique (*CYCLIN DEPENDENT KINASE D1*, score -8,91 ; *POLTERGEIST*, score -2,39 ; *SHOOT MERISTEMLESS*, score -8,2), dans les remaniements de la paroi cellulaire (*CELLULOSE SYNTHASE 6*, score -9,78 ; *EXPANSINE B3*, score -9,65) ou encore dans la signalisation auxinique (*PIN7*, score -5,49 ; *PIN4*, score -8,36 ; *IAA7*, score -8,65) sont associés à des sites de liaison avec un haut score pour PpLFY1, et quelques termes de Gene Ontology associés à ces mêmes fonctions sont surreprésentés. Ces résultats sont bien sûr à nuancer puisque les relations d'orthologie entre les gènes d'*A. thaliana* et de *P. patens* sont difficiles à établir, étant donné la grande distance évolutive entre ces deux espèces. Dire qu'un noyau de gènes est régulé communément par LFY et PpLFY n'aura pas forcément beaucoup de sens puisque ces gènes ont énormément évolué entre les deux espèces. Des études sur la dérégulation des gènes chez un mutant *lfy* de *Marchantia polymorpha*, qui appartient au groupe des mousses et se positionne donc plus proche de *P. patens*, sont en cours dans l'équipe de Takashi Araki, et pourraient se révéler plus instructives pour la comparaison avec les gènes dérégulés chez *P. patens*.

La fonction de certains des gènes prédits par notre modèle, en cohérence avec le phénotype mutant *pplfy1 pplfy2*, suggère que *PpLFY1/2* contrôlent des aspects reliés à la division et à l'expansion cellulaire. Le phénotype des plantes *pAct::PpLFY1* est en accord avec cette hypothèse, puisque ces surexpresseurs développent des caulonémas courts et peu ramifiés. Le phénotype drastique des plantes *pHSP::VP16-PpLFY1* après induction par choc thermique, ainsi que le fait que peu de surexpresseurs forts *pAct::PpLFY1* aient pu être isolés,

suggère qu'un niveau « adéquat » d'expression de *PpLFY1* est essentiel au développement normal de la plante, ce qui est compréhensible si *PpLFY1* a un rôle dans le contrôle de la division cellulaire. Il faut néanmoins préciser que la surexpression d'une construction *35S::VP16-LFY* chez *A. thaliana* entraîne également un phénotype très marqué (mort de la plante après jaunissement rapide, communication personnelle de Detlef Weigel), et il est donc possible que la surexpression de nombreux orthologues de *LFY* fusionnés au domaine VP16 entraîne un phénotype comparable, et pas seulement dans le cas de *PpLFY1*. Comme proposé dans l'Article 1, nous pensons que le contrôle de la division cellulaire et de l'état méristématique par PpLFY reflète une fonction ancestrale de LFY, qui est encore directement visible chez *P. patens*, et qui se devine chez certaines espèces d'angiospermes. En effet chez le pois, l'orthologue de *LFY*, *UNI*, est exprimé dans la bordure des feuilles composées, et a un rôle dans l'initiation des folioles par le maintien d'un état indifférencié transitoire à cette bordure (Hofer et al., 1997). Le même phénomène est observé chez le riz, puisque *RFL* est exprimé dans les méristèmes de ramifications de l'inflorescence où il semble maintenir un état d'indifférenciation nécessaire à l'émergence des ramifications (Kyozuka et al., 1998; Rao et al., 2008). Chez *A. thaliana*, bien que la fonction de *LFY* dans le contrôle de l'état méristématique n'apparaisse pas de façon évidente, elle est sans doute redondante avec d'autres facteurs puisque le triple mutant *pny pnf /+ lfy (PNY, PENNYWISE ; PNF, POUNDFOOLISH)*, par exemple, présente une forte réduction du nombre de méristèmes axillaires (Kanrar et al., 2008). De plus, de nombreux gènes identifiés en ChIP-Seq, qui ne semblent pas reliés à la fonction florale de LFY, témoignent peut-être de cette fonction méristématique ancestrale.

La découverte d'orthologues de *LFY* chez certaines algues vertes relance la question de sa fonction ancestrale. A-t-elle aussi un lien avec la division cellulaire et l'état d'indifférenciation des structures cellulaires chez ces espèces ? Le fait que *LFY* apparaisse en même temps que la multicellularité pourrait signifier qu'il a été nécessaire à cet évènement, ce qui est envisageable si *LFY* contrôle des aspects de la division cellulaire. Aucun génome d'algue verte multicellulaire n'est disponible pour le moment ; nous ne pourrons donc pas rechercher des gènes cibles putatifs de LFY grâce à des outils prédictifs. Le génome le plus proche qui soit disponible est celui de *Chlamydomonas reinhardtii*, algue verte unicellulaire modèle où des outils de transgénèse sont disponibles ; il pourrait donc être envisagé de rechercher les sites de liaison de LFY dans ce génome et d'analyser les conséquences d'une surexpression de *LFY* chez cette espèce, bien qu'un tel système hétérologue soit difficile à

interpréter. Nous sommes donc encore assez loin de pouvoir valider le rôle ancestral de LFY chez les algues, mais les progrès très rapides en séquençage haut-débit nous fourniront vraisemblablement bientôt des données génomiques à analyser chez des algues vertes multicellulaires. Tout ceci nous permettra peut-être de comprendre quelle est l'origine de LFY et de son réseau transcriptionnel, et pourquoi ce réseau a évolué vers le contrôle du développement floral chez les angiospermes, groupe au succès évolutif majeur.

En conclusion, grâce à l'étude de la spécificité de liaison de LFY chez la lignée verte, nous avons pu développer un modèle de prédiction de liaison de LFY à l'ADN et suivre l'histoire de son réseau transcriptionnel, malgré la variation des propriétés (séquence, score de liaison, position, nombre) de ses éléments *cis*. De plus, contrairement à ce qui était attendu, cette étude a démontré qu'un facteur de transcription sans famille multigénique et au rôle essentiel a pu changer de spécificité de liaison plusieurs fois au cours de l'évolution. Nous avons donc découvert que le réseau contrôlé par LFY montrait une grande fluidité en *cis*, et quelques évènements plus ponctuels mais aussi beaucoup plus drastiques de changements en *trans*. LFY constitue un modèle de choix pour comprendre l'évolution des facteurs de transcription, et nous pensons que ces résultats peuvent avoir une portée très générale. Etendre ces études inter-espèces à de nombreux autres facteurs de transcription nous révèlera si de telles conclusions peuvent être généralisées à l'évolution de tous les réseaux transcriptionnels.

MATERIEL ET METHODES

Les Matériel et Méthodes des études publiées ou en cours de soumission sont détaillés dans les articles correspondants.

I) Détermination de la spécificité de liaison à l'ADN de LFY : des clonages au SELEX

Clonages et origine des différents ADNc utilisés :
Les plasmides contenant les ADNc de *RFL* (*Oryza sativa*), *RoLFY* (*Rosa chinensis*), *VFL* (*Vitis vinifera*), *GbLFY* et *GbNLY* (*Ginkgo biloba*), *WelLFY* et *WelNLY* (*Welwitschia mirabilis*), *PaNLY* (*Picea abies*), *CrLFY2* (*Ceratopteris richardii*) et *PpLFY1* (*Physcomitrella patens*), et *MarpoFLO* (*Marchantia polymorpha*) ont été fournis par les équipes de Junko Kyozuka, Mohammed Bendahmane, Jean Masson, Michael Frohlich, Peter Engstrom et Mitsuyasu Hasebe que nous remercions vivement. La séquence d'*AmboLFY* (*Amborella trichopoda*) a été isolée par Edwige Moyroud. Les fragments d'intérêt sont amplifiés par PCR grâce à la Taq Phusion (Thermo Scientific), en rajoutant des sites de restriction aux extrémités des produits. Les amplicons sont purifiés sur gel d'agarose (kit BioBasic Inc.) puis insérés dans le plasmide pCR-Blunt (Invitrogen). Après validation de leur intégrité par séquençage, les inserts sont transférés dans les plasmides pETM-11 (Dummler et al., 2005) ou pET30a+ (Novagen) grâce aux enzymes de restriction appropriées. Toutes les étapes de transformation sont réalisées en bactéries *Escherichia coli* thermocompétentes de la souche DH5α, avec 10 min de contact entre les bactéries et le plasmide sur glace, 1 min de choc thermique à 42°C, ajout de 800 µL de milieu LB (Luria Broth), puis croissance pendant 45 minutes à 37°C avec agitation. Les bactéries sont ensuite étalées sur milieu LB - agar contenant de la kanamycine (50 mg/mL) pour sélectionner les transformants.

Purification de protéines recombinantes :
Les plasmides pETM-11 ou pET30a+ contenant les ADNc d'intérêt sont transformés en bactéries *Escherichia coli* de la souche Rosetta2, optimisée pour la production de protéines recombinantes. Après transformation, les bactéries sont étalées sur milieu LB - agar contenant de la kanamycine (50 mg/mL), la résistance étant apportée par le plasmide pETM-11 ou pET30a+, et du chloramphénicol (34 mg/mL), la résistance étant apportée par le plasmide pRARE de la souche bactérienne. Le lendemain, 10 mL de LB liquide (avec kanamycine et chloramphénicol) sont ensemencés avec les colonies obtenues, à 37°C avec agitation. 45 min plus tard, 5 mL de milieu sont rajoutés ; cette étape est répétée une seconde fois. Enfin, la culture est transférée dans 1L de LB + kanamycine + chloramphénicol. Une fois la culture saturée, de la bétaïne à 2mM est rajoutée à la culture et l'induction de l'expression de la protéine est réalisée à 17°C avec 0,4 mM d'IPTG. Pour RFL-C, LFY-C, PaNLY-C et RoLFYΔ, la culture a été effectuée en milieu LB + NaCl 0,5 M pour améliorer l'induction de l'expression de la protéine. Après 12 à 16 h d'induction, la culture est centrifugée à 4°C à 4500 rpm pendant 30 minutes. Le culot est lavé avec une solution de Tris 100 mM pH 8, puis centrifugé à nouveau pendant 20 minutes à 4000 rpm à 4°C. Le culot est repris dans le tampon de sonication (voir Tableau) additionné d'une tablette d'anti-protéases (CIP complete EDTA free, Roche). La sonication est effectuée dans un mélange eau-éthanol pendant 12 minutes, à fréquence de pulsations 40%, et puissance 6,5. L'extrait brut est centrifugé pendant 40 minutes à 16000 rpm à 4°C pour récupérer la fraction soluble, qui est reprise dans le tampon de base. La fraction soluble est déposée sur une colonne de résine Nickel-Sépharose (GE Healthcare), préalablement équilibrée avec le tampon d'équilibration. 20 mL de solution de lavage sont appliqués sur la colonne, puis 10 mL de solution d'élution, et la protéine est récupérée en fractions d'élution de 1,5 mL. La protéine est dialysée sur la nuit pour éliminer l'imidazole. Si nécessaire, la protéine est concentrée sur une colonne Vivaspin (Sartorius Vivascience) par centrifugations successives de 5-10 min à 3000 rpm à 4°C. Une

purification par exclusion de taille a été effectuée pour RFL-C et VFLΔ, avec une colonne AKTA HiLoad 16/60 Superdex 200, en tampon Tris 20 mM pH 8, NaCl 200 mM, DTT 5 mM.

Protéines concernées	Tampon de sonication	Tampon de base	Solution d'équilibration	Solution de lavage	Solution d'élution	Tampon de dialyse
RFL-C, RoLFY-C, VFL-C, MarpoFLO-C, PaNLY-C	Tris 20 mM pH 8 NaCl 250 mM glycérol 5% TCEP 5 mM	Tris 20 mM pH 8 NaCl 250 mM glycérol 5% DTT 1 mM	Tampon de base + 5 mM imidazole	Tampon de base + 50 mM imidazole	Tampon de base + 350 mM imidazole	Tris 20 mM pH 7,5 NaCl 150 mM glycérol 5% EDTA 0,25 mM $MgCl_2$ 2 mM DTT 5 mM
RFLΔ, RoLFYΔ, VFLΔ, CrLFY2Δ	Tris 50 mM pH 8 TCEP 3 mM	Tris 20 mM pH 7,5 TCEP 1 mM	Tampon de base + 10 mM imidazole	Tampon de base + 20 mM imidazole	Tampon de base + 300 mM imidazole	Tris 20 mM pH 7,5 glycérol 5% DTT 5 mM
PpLFY1, LFYΔ, et protéines mutées associées (Fig. 21)	Tris 50 mM pH 8 TCEP 3 mM $MgCl_2$ 1 mM glycérol 10%	Tris 25 mM pH 8 TCEP 1 mM	Tampon de base + 5 mM imidazole	Tampon de base + 20 mM imidazole	Tampon de base + 300 mM imidazole	Tris 25 mM pH 8 EDTA 0,5 mM DTT 1 mM

Bilan des tampons utilisés pour les purifications des différentes protéines de cette étude.

Gel retard (EMSA : Electrophoretic Mobility Shift Assay):

Les oligonucléotides fluorescents sont obtenus par commande de l'oligo sens marqué en TAMRA en 5', ou bien par ajout d'un nucléotide fluorescent en 5' : pour cela, les oligonucléotides sont dessinés pour obtenir un G dépassant en 5' du brin sens. Les brins sens et antisens sont mélangés à 10 µM chacun dans un tampon d'hybridation (Tris 10 mM pH 7,5, NaCl 150 mM, EDTA 1 mM), chauffés à 95°C pendant 5 minutes, et refroidis jusqu'à température ambiante pendant 3h. Pour le marquage, 20 µL d'oligonucléotides double brin à 200 nM sont mélangés à 1 µL de Cy5-dCTP ou de Cy3-dCTP à 8 µM (GE Healthcare) et 1 Unité d'enzyme Klenow (New England Biolabs) pendant 1h à 37°C à l'obscurité, suivi d'une inactivation de l'enzyme pendant 10 min à 65°C. Pour l'EMSA, le mélange ADN-protéine est réalisé dans le tampon Binding Buffer (Tris 20 mM pH 7,5, NaCl 150 mm, EDTA 0,25 mM, $MgCl_2$ 2 mM, glycérol 1 %, TCEP 3 mM) pour les constructions LFY C-terminales, et dans le tampon « Eugenio » (Hepes 10 mM pH 7,2, Spermidine 1 mM, EDTA 14 mM, BSA 0,3 mg/mL, CHAPS 0,25 %, glycérol 1 %, TCEP 3 mM) pour les constructions Δ ou entières. Pour chaque réaction de 20 µL, on utilise 10 nM d'ADN, de l'ADN-Fish (ADN de sperme de poissons (Roche)) comme compétiteur à 28 ng/mL, et une concentration variable de protéine (de l'ordre de 50 nM à 5 µM). La réaction reste sur glace pendant 15 min. Le gel d'acrylamide (6 % ou 8 %) subit une première migration dans du TBE 0,5X pendant 30 minutes à 90 V et à 4°C, avant dépôt du mélange ADN-protéine. La migration est effectuée pendant 1h10 (pour un gel à 6 %) ou 1h30 (pour un gel à 8 %), à 90 V et à 4°C, à l'obscurité. Le gel après migration est scanné grâce au Typhoon 9400 (Molecular Dynamics), avec une longueur d'onde d'excitation de 580 nm et 670 nm et des valeurs de PMT (PhotoMultiplier Tube) de 550 et 530 respectivement pour le Cy3 et le Cy5.

SELEX (Systematic Enrichment of Ligands by EXponential enrichment) :

Pour créer l'oligonucléotide "Random" double brin initial, l'oligonucléotide Random-Matrix ($TGGAGAAGAGGAGAGATCTAGC(N)_{30}CTTGTTCTTCTTCGATTCCGG$) est amplifié par PCR (25 cycles avec 10'' de dénaturation à 98°C, 30'' d'hybridation à 55°C et 20'' d'amplification à 72°C), avec les amorces SELEX-F (TGGAGAAGAGGAGAGATCTAG, marquée au TAMRA en 5') et SELEX-R (CCGGAATCGAAGAAGAACAA), avec la Taq Polymérase Phusion (New England Biolabs). L'ADN amplifié est dosé sur gel d'acrylamide 6% face à une gamme, et dilué à 100nM. La protéine (500 nM ou 1 µM) est mise en contact avec 22,5 µL de l'ADN « Random », le mélange est complété à 225 µL avec le tampon SELEX (Tris 20 mM pH 8, NaCl 50 mM, $MgCl_2$ 2 mM, glycérol 1%) et de l'ADN-Fish (60 µg/mL), et mis à tourner sur roue à 4°C pendant 15 minutes. Les billes de nickel magnétiques (Ni-NTA magnetic agarose beads, Qiagen, 25 µL par réaction) sont lavées deux fois avec le tampon SELEX : le tampon est rajouté aux billes, le tube est posé sur un aimant et le liquide est enlevé. Les billes lavées sont ensuite rajoutées au mélange ADN-protéine, qui est mis à tourner sur roue à 4°C pendant 30 minutes. Le tube est ensuite placé sur l'aimant pour évacuer le liquide, et les billes subissent 6 séries de lavages avec 50 µL de tampon SELEX puis 250 µL de tampon SELEX + ADN-Fish (20 µg/mL). 2 µL de billes sont prélevées aux lavages 0, 2, 4 et 6. Les oligonucléotides sélectionnés sont amplifiés par PCR avec la Taq polymérase Phusion sur 1 µL de

billes (20 cycles avec 10'' de dénaturation à 98°C, 25'' d'hybridation à 55°C et 15'' d'amplification à 72°C), avec les amorces SELEX-F et SELEX-R et la Taq Phusion. Les produits PCR sont dosés sur gel d'acrylamide, puis 2 µL d'ADN à 100 nM sont testés en EMSA avec la protéine. Selon le résultat de l'EMSA, les produits PCR issus des lavages 4 ou 6 sont mis en contact avec la protéine pour effectuer le cycle de sélection suivant. Une fois qu'un enrichissement suffisant est obtenu, les billes correspondantes sont à nouveau amplifiées par PCR avec les amorces SELEX-F non marquée et SELEX-R. 1 µL de produit PCR est ligué dans le plasmide pCR-Blunt pendant 1h à 16°C, puis la réaction de ligation est transformée en bactéries DH5α. Une PCR sur colonie est réalisée sur une dizaine de colonies, avec les amorces SELEX-F (marquée au TAMRA) et SELEX-R. 2 µL de produit PCR sont testés en EMSA avec la protéine. Les colonies qui montrent un retard sont envoyées à séquencer. Le SELEX est considéré comme terminé si toutes les colonies ou presque ont montré un retard de migration, et si les séquences associées sont variées, témoignant d'un enrichissement adéquat.

Modification des oligonucléotides de SELEX pour le séquençage Illumina:

Les oligonucléotides sélectionnés par le SELEX sont amplifiés avec les amorces AdF-x (ACACTCTTTCCCTACACGACGCTCTTCCGATC$(N)_6$TTGGAGAAGAGGAGAGATCTAGC) et Ad-R (CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTCCGGAATCGAAGAAGAACAAG) pour rajouter le code-barre numéro x (de séquence $(N)_6$), la zone d'amorce pour le séquençage, et l'une des séquences d'hybridation à la puce en 3'. Une deuxième réaction PCR avec les amorces PCR-F (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT) et PCR-R (CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT) permet de rajouter l'autre séquence d'hybridation à la puce en 5'. Toutes les PCR sont effectués avec la Taq polymerase Platinium Pfx (Invitrogen). Les fragments amplifiés sont purifiés par extraction sur gel d'agarose 3%, puis évalués pour leur quantité et pureté grâce au BioAnalyser 2100 et au kit DNA 1000 (Agilent). Les échantillons sont mélangés ensemble à 10 nM chacun, et séquencés sur une plateforme Illumina HiSeq 2000 avec un programme de type 70-bp single end, à 10 % de la capacité totale d'une puce de séquençage.

Traitement bioinformatique des séquences de SELEX:

Les séquences ont été filtrées selon leur qualité par le logiciel SHORE (http://1001genomes.org/software/shore.html) par Norman Warthmann. Les code-barres et les bordures constantes ont été identifiés par des programmes créés en Python (version 2.6, et librairie NumPy). Pour cela, un code-barre strict est recherché, mais les bordures peuvent varier jusqu'à 6 pb à chaque extrémité. Les séquences sont triées par nombre d'occurrence et les séquences répétées sont éliminées. Les 2000 premières séquences sont alignées grâce à MEME, en recherchant un motif palindromique de 19 pb pour toutes les protéines LFY, et sans contraintes pour les protéines NLY. Les matrices de fréquences issues de l'alignement de MEME ont été soumises pour comparaison par le logiciel STAMP (http://www.benoslab.pitt.edu/stamp/index.php) avec les paramètres par défaut.

Expériences de QuMFRA (Quantitative Multiple Fluorescence Relative Affinity) :

Un gel retard est réalisé comme précédemment, et dans chaque piste un oligonucléotide de référence marqué au fluorophore Cy3, et un oligonucléotide à tester marqué au fluorophore Cy5, sont incubés avec la protéine. L'intensité des bandes d'ADN libre ou complexé à la protéine est mesurée avec le logiciel ImageQuant, et le K_D relatif de l'oligonucléotide à tester est calculé comme décrit dans (Man and Stormo, 2001). Chaque mesure est répétée au moins 2 fois. Le score relatif expérimental est :

$$S = -\ln\left(\frac{K_D}{K_{D,\min}}\right)$$

où $K_{D,min}$ est le K_D relatif le plus faible mesuré. On peut déduire de la corrélation entre le score prédit S et le K_D relatif expérimental les valeurs a et b de l'équation suivante (équation 1), nécessaires au calcul de l'Occupation par la suite :

$$S = -a \times \ln(K_D) + b$$

Optimisation de la matrice de PpLFY1 :
51 oligonucléotides ont été testés en QuMFRA, et pour 34 d'entre eux le K_D relatif a été déterminé. Le score de ces oligonucléotides a été calculé avec la matrice issue de l'alignement de 514 séquences du SELEX de PpLFY1 dans un premier temps (matrice symétrique). La fréquence des doublets aux positions 2-18 et 3-17 a été calculée à partir de ces 514 séquences alignées (matrice symétrique + doublets (1)), ou bien à partir des 49 379 séquences uniques initiales (matrice symétrique + doublets (2)). Pour cela, le score de tous les sites de liaison de 19 pb de chacune des 49 379 séquences uniques a été calculé avec la matrice symétrique, pour garder à chaque fois la position correspondant au score maximum. Une fois cet « alignement » déterminé, la fréquence des doublets a été calculée.

II) Prédire la liaison de LFY à l'ADN

Calcul de score et d'Occupation Prédite (POcc) :
Tous les programmes informatiques ont été écrits en langage Python (www.python.org), en utilisant la libraire numérique NumPy. Le calcul de score est décrit dans la figure 14. Les programmes de calcul de POcc ont été écrits par Eugenio Gomez-Minguet. L'Occupation Prédite, définie comme le nombre de molécules du facteur de transcription (TF) liées sur un site s de longueur W sur une séquence entière de longueur L, est calculée comme suit (Roider et al., 2007) :

$$POcc = \sum_{1}^{L-W} \frac{K_A \times [TF]}{1 + K_A \times [TF]}$$

où K_A est la constante d'association de la réaction de liaison de TF au site. Cette constante est l'inverse de la constante de dissociation K_D qui peut être calculée grâce à l'équation 1, le score S étant calculé comme expliqué précédemment avec la matrice poids/position de la protéine considérée, et a et b étant déterminés de manière expérimentale par QuMFRA (ou bien par défaut a=1 et b=0). Par défaut, on pose [TF] égal à exp(b/a).

Analyse de la conservation des sites de liaison de LFY entre *A. thaliana* et *A. lyrata* :
La valeur de POcc de toutes les séquences de 150 pb glissantes du génome d'*A. thaliana* a été calculée, puis les régions adjacentes possédant une valeur de POcc supérieure au seuil choisi sont fusionnés. Pour chercher l'orthologue d'une région R chez *A. lyrata*, l'orthologue du gène le plus proche de R a d'abord été recherché, les relations d'orthologie entre gènes d'*A. lyrata* et *A. thaliana* ayant été fournies par Roberto Solano. La distance entre la région R et le gène (CDS) le plus proche chez *A. thaliana* est D ; par analogie l'orthologue de la région R est situé à la distance D de l'orthologue du gène associé chez *A. lyrata.* Une fenêtre de 1000 pb autour de cette région est sélectionnée, et la POcc est calculée à nouveau sur les 150 pb glissantes de ce fragment. La séquence de 150 pb montrant la plus haute valeur de POcc est considérée orthologue à la région R. La région est considérée comme liée en ChIP-Seq si au moins 1 nucléotide est situé sous le pic de ChIP-Seq. Les données de micro-array utilisées pour analyser la dérégulation des gènes associés aux régions à forte Occ sont issues de (Schmid et al., 2003; Schmid et al., 2005), et ont été récupérées par le serveur Gene Expression Omnibus pour la dérégulation lors de la transition florale (array GDS453) et par le site AtGenExpress pour la dérégulation chez *lfy-12*.

Prédiction des sites de liaison de PpLFY1 au génome de *P. patens* :
La valeur de POcc de toutes les séquences de 150 pb glissantes du génome de *P. patens* a été calculée. Le gène le plus proche d'une région à forte valeur de POcc est celui dont le site de démarrage ou de terminaison de la transcription est le plus proche de la région. La position des sites de démarrages ou de terminaison de la transcription des gènes de *Physcomitrella patens* ainsi que leur relation d'orthologie avec les gènes d'*Arabidopsis thaliana* sont issus des fichiers d'annotation du génome, disponibles sur le site Phytozome (http://www.phytozome.net/). Les termes de Gene Ontology associés aux gènes d'*Arabidopsis thaliana* sont issus des fichiers de données brutes de TAIR (http://www.arabidopsis.org/).

III) Culture de *Physcomitrella patens*, et méthodes de biologie moléculaire associées

Culture de *Physcomitrella patens* :
La souche sauvage de *P. patens* a été fournie par Fabien Nogué et Florence Charlot (Institut Jean Pierre Bourgin, Versailles). Les souches sont cultivées sur milieu minimal $PpNO_3$ ($Ca(NO_3)_2$,$4H_2O$ 0,8 g/L, $MgSO_4$,$7H_2O$ 0,25 g/L, $FeSO_4$,$7H_2O$ 0,0125 g/L, KH_2PO_4 25 µg/L, micro-éléments) parfois additionné de tartrate d'ammonium ($(NH_4)_2C_4H_4O_6$, 500 mg/L). De la céfotaxime (Biochemicals Direct) à 150 mg/L est parfois additionnée pour limiter les contaminations bactériennes. Une feuille de cellulose (AA Packagings Limited) stérile est déposée sur le milieu, et les clones sont mis à pousser dessus. Pour régénérer les cultures, un échantillon de gamétophyte est broyé dans de l'eau stérile, grâce à un homogénéisateur Ultra-Turrax T18 (IKA), puis réétalé sur boîte. Les plantes sont cultivées à 20°C en jours longs (16h de jour, 8h de nuit). Les clones sont observés avec une loupe binoculaire Olympus SZX12.

Clonages pour surexprimer *PpLFY1*:
L'ADNc de *PpLFY1* a été fourni par Mitsuyasu Hasebe. Les constructions comprenant les promoteurs pHSP ou pAct ont été fournies par Fabien Nogué. Ces constructions comprennent une région d'homologie avec le locus 108 de *P. patens* (Schaefer, 2001), fréquemment utilisé pour des expériences de transformation. Au moins 50 µg des plasmides obtenus après clonage sont digérés pour libérer la zone contenant le locus 108 de part et d'autre, et la construction à insérer au centre. La digestion est arrêtée par inactivation des enzymes à 80°C pendant 20 minutes, puis l'ADN est précipité à l'éthanol (2,5 volumes) et à l'acétate de sodium (0,1 volumes) pendant 1h à -20°C. Après centrifugation, l'ADN est lavé avec 2 volumes d'éthanol 70%, puis centrifugé à nouveau. Après séchage, le culot est repris dans 50 µL d'eau stérile et dosé au NanoDrop.

Transformation de *Physcomitrella patens* :
Des protonémas de 7 jours sont incubés avec 10 mL d'une solution de Drisélase 2 % (Sigma) et mannitol 8,5 % pendant 30 minutes à température ambiante. Les protoplastes sont filtrés sur tamis de 80 µm puis 40 µm, centrifugés à 600 g pendant 5 minutes et lavés deux fois avec 10 mL de mannitol 8,5 %. La densité de protoplastes est mesurée sur cellule de Malassez au microscope optique, pour diluer la solution à 1,2 x 10^6 protoplastes / mL dans du milieu MMM (mannitol 8,5 %, $MgCl_2$ 0,305 %, MES 0,1 %). 300 µL de protoplastes sont mélangés à 10-15 µg de plasmide linéarisé par une digestion enzymatique appropriée, et à 300 µL d'une solution stérile de PEG 4000 (Polyéthylène glycols). L'ensemble subit un choc thermique de 5 minutes à 45°C, puis est laissé à reposer pendant 10 minutes à température ambiante. 6 mL de milieu $PpNH_4$ + mannitol 8,5% sont rajoutés progressivement aux protoplastes, qui sont laissés ensuite sur la nuit à l'obscurité en chambre de culture. Ils sont ensuite coulés avec du milieu Top Layer (agar 14 g/L, mannitol 8,5%), étalés sur milieu $PpNH_4$ + mannitol 8,5%, et remis en chambre de culture à la lumière. Quelques jours plus tard, les clones sont transférés sur milieu sélectif pour une semaine, transférés sur milieu sans antibiotique pour une semaine, et replacés sur milieu sélectif pour une semaine. Les clones résistants après la deuxième étape de sélection sont considérés comme stables.

RT-PCR (Reverse Transcription suivie d'une PCR):
Les ARNs sont extraits de clones au stade gamétophytique grâce au kit RNeasy Mini (Qiagen). Après quantification par Nanodrop, 1 µg d'ARN sont incubés avec 2 µL d'oligo dT (1 µg/µL) pendant 5 minutes à 70°C, puis laissés à reposer 5 minutes dans la glace. Les ADNc sont synthétisés avec 200 unités de Reverse Transcriptase M-MLV (Promega), 1 µL d'inhibiteur de ribonucléase (RNasin, Promega), 0,5 mM de dNTPs et le buffer M-MLV 1X, pendant 1h à 40°C, suivi de 15 minutes à 70°C. La PCR semi-quantitative pour *PpLFY1* est effectuée sur 1 µL d'ADNc avec les amorces PpLFY1-F (5'-GGAGCAACAGCGCATGGATTTG) et PpLFY1-R (5'-CACCAACATTTTCTCCAC GCTCTTT), avec 25 cycles d'amplification et la Taq Polymérase Phusion (New England Biolabs). L'expression de *VP16-PpLFY1* est mesurée avec les amorces pHSP-F (5'-

GTAGATTCAACCTCAATTTGCAGAG) située dans le promoteur *pHSP*, et PpLFY1-R, avec 30 cycles d'amplification. Le gène de référence *PpAPT* est amplifié avec les amorces PpAPT#16 et PpAPT#19 (5'-CCACCCATTGCTCTTGCCATC et 5'-CCCGACAACTTCTCACGACCC) pendant 25 cycles. La quantification des bandes est réalisée sur ImageJ. Pour les clones *pAct::PpLFY1*, l'expression de *PpLFY1* a été mesurée 3 fois à partir des mêmes échantillons d'ARN. Pour les clones *pHSP::VP16-PpLFY1*, l'expression de *VP16-PpLFY1* n'a été mesurée qu'une seule fois pour le moment.

LFY chez les algues vertes :

Les séquences des ADNc de LFY de *Nothoceros aenigmaticus*, *Coleochaete scutata*, *Cylindrocystis sp.* et *Klebsormidium subtile* ont été obtenues grâce à une collaboration avec Edwige Moyroud et Samuel Brockington (University of Cambridge). Les phylogénies des figures 25 et 26 ont été effectuées sur le site Phylogeny.fr, avec l'algorithme d'alignement Muscle, un raffinement par GBlocks (Fig. 25B et Fig. 26B) ou Built-in-Curer (Fig. 25A), et une phylogénie construite avec PhyML.

REFERENCES BIBLIOGRAPHIQUES

Aagaard, J.E., Olmstead, R.G., Willis, J.H., and Phillips, P.C. (2005). Duplication of floral regulatory genes in the Lamiales. Am J Bot **92,** 1284-1293.

Adams, K.L., and Wendel, J.F. (2005). Polyploidy and genome evolution in plants. Curr Opin Plant Biol **8,** 135-141.

Airoldi, C.A., and Davies, B. (2012). Gene duplication and the evolution of plant MADS-box transcription factors. J Genet Genomics **39,** 157-165.

An, X.M., Wang, D.M., Wang, Z.L., Li, B., Bo, W.H., Cao, G.L., and Zhang, Z.Y. (2011). Isolation of a LEAFY homolog from Populus tomentosa: expression of PtLFY in P. tomentosa floral buds and PtLFY-IR-mediated gene silencing in tobacco (Nicotiana tabacum). Plant Cell Rep **30,** 89-100.

Arthur, W. (2002). The emerging conceptual framework of evolutionary developmental biology. Nature **415,** 757-764.

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol **2,** 28-36.

Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola, T.S., Young, R.A., and Gifford, D.K. (2003). Computational discovery of gene modules and regulatory networks. Nat Biotechnol **21,** 1337-1342.

Barton, M.K., and Poethig, R.S. (1993). The formation of the shoot apical meristem in *Arabidopsis thaliana*: an analysis of development in the wild type and in the *shoot meristemless* mutant. Development **119,** 823-831.

Baum, D.A., Yoon, H.S., and Oldham, R.L. (2005). Molecular evolution of the transcription factor LEAFY in Brassicaceae. Mol Phylogenet Evol **37,** 1-14.

Benlloch, R., Berbel, A., Serrano-Mislata, A., and Madueno, F. (2007). Floral initiation and inflorescence architecture: a comparative view. Ann Bot **100,** 659-676.

Berleth, T., and Jürgens, G. (1993). The role of the *monopteros* gene in organising the basal body region of the *Arabidopsis* embryo. Development **118,** 575-587.

Blazquez, M.A., and Weigel, D. (2000). Integration of floral inductive signals in Arabidopsis. Nature **404,** 889-892.

Blazquez, M.A., Soowal, L.N., Lee, I., and Weigel, D. (1997). LEAFY expression and flower initiation in Arabidopsis. Development **124,** 3835-3844.

Bomblies, K., Wang, R.L., Ambrose, B.A., Schmidt, R.J., Meeley, R.B., and Doebley, J. (2003). Duplicate FLORICAULA/LEAFY homologs zfl1 and zfl2 control inflorescence architecture and flower patterning in maize. Development **130,** 2385-2395.

Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R., Wang, L.Y., Gerstein, M., and Snyder, M. (2007). Divergence of transcription factor binding sites across related yeast species. Science **317,** 815-819.

Bradley, D., Carpenter, R., Sommer, H., Hartley, N., and Coen, E. (1993). Complementary floral homeotic phenotypes result from opposite orientations of a transposon at the *plena* locus of Antirrhinum. Cell **72,** 85-95.

Busch, M.A., Bomblies, K., and Weigel, D. (1999). Activation of a floral homeotic gene in Arabidopsis. Science **285,** 585-587.

Carmona, M.J., Cubas, P., and Martinez-Zapater, J.M. (2002). VFL, the grapevine FLORICAULA/LEAFY ortholog, is expressed in meristematic regions independently of their fate. Plant Physiol **130,** 68-77.

Causier, B., Bradley, D., Cook, H., and Davies, B. (2009). Conserved intragenic elements were critical for the evolution of the floral C-function. Plant J **58,** 41-52.

Causier, B., Castillo, R., Zhou, J., Ingram, R., Xue, Y., Schwarz-Sommer, Z., and Davies, B. (2005). Evolution in action: following function in duplicated floral homeotic genes. Curr Biol **15,** 1508-1512.

Chae, E., Tan, Q.K., Hill, T.A., and Irish, V.F. (2008). An Arabidopsis F-box protein acts as a transcriptional co-factor to regulate floral development. Development **135,** 1235-1245.

Chang, Y., and Graham, S.W. (2011). Inferring the higher-order phylogeny of mosses (Bryophyta) and relatives using a large, multigene plastid data set. Am J Bot **98,** 839-849.

Chung, H., Bogwitz, M.R., McCart, C., Andrianopoulos, A., Ffrench-Constant, R.H., Batterham, P., and Daborn, P.J. (2007). Cis-regulatory elements in the Accord retrotransposon result in tissue-specific expression of the Drosophila melanogaster insecticide resistance gene Cyp6g1. Genetics **175,** 1071-1077.

Clarke, J.T., Warnock, R.C., and Donoghue, P.C. (2011). Establishing a time-scale for plant evolution. New Phytol **192,** 266-301.

Clarke, N.D., and Granek, J.A. (2003). Rank order metrics for quantifying the association of sequence features with gene regulation. Bioinformatics **19,** 212-218.

Coen, E.S., and Meyerowitz, E.M. (1991). The war of the whorls: genetic interactions controlling flower development. Nature **353,** 31-37.

Coen, E.S., Romero, J.M., Doyle, S., Elliott, R., Murphy, G., and Carpenter, R. (1990). floricaula: a homeotic gene required for flower development in antirrhinum majus. Cell **63,** 1311-1322.

Cohn, M.J., and Tickle, C. (1999). Developmental basis of limblessness and axial patterning in snakes. Nature **399,** 474-479.

Cove, D. (2005). The moss Physcomitrella patens. Annu Rev Genet **39,** 339-358.

Daborn, P.J., Yen, J.L., Bogwitz, M.R., Le Goff, G., Feil, E., Jeffers, S., Tijet, N., Perry, T., Heckel, D., Batterham, P., Feyereisen, R., Wilson, T.G., and ffrench-Constant, R.H. (2002). A single p450 allele associated with insecticide resistance in Drosophila. Science **297,** 2253-2256.

Davies, B., Motte, P., Keck, E., Saedler, H., Sommer, H., and Schwarz-Sommer, Z. (1999). PLENA and FARINELLI: redundancy and regulatory interactions between two Antirrhinum MADS-box factors controlling flower development. EMBO J **18,** 4023-4034.

Djordjevic, M. (2007). SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways. Biomol Eng **24,** 179-189.

Dowell, R.D. (2010). Transcription factor binding variation in the evolution of gene regulation. Trends Genet **26,** 468-475.

Dummler, A., Lawrence, A.M., and de Marco, A. (2005). Simplified screening for the detection of soluble fusion constructs expressed in E. coli using a modular set of vectors. Microb. Cell Fact. **4,** 34.

Ferrandiz, C., Gu, Q., Martienssen, R., and Yanofsky, M.F. (2000). Redundant regulation of meristem identity and plant architecture by FRUITFULL, APETALA1 and CAULIFLOWER. Development **127,** 725-734.

Friis, E.M., Pedersen, K.R., and Crane, P.R. (2005). When Earth started blooming: insights from the fossil record. Curr Opin Plant Biol **8,** 5-12.

Frohlich, M.W. (2003). An evolutionary scenario for the origin of flowers. Nat Rev Genet **4,** 559-566.

Frohlich, M.W., and Chase, M.W. (2007). After a dozen years of progress the origin of angiosperms is still a great mystery. Nature **450,** 1184-1189.

Godoy, M., Franco-Zorrilla, J.M., Perez-Perez, J., Oliveros, J.C., Lorenzo, O., and Solano, R. (2011). Improved protein-binding microarrays for the identification of DNA-binding specificities of transcription factors. Plant J **66,** 700-711.

Granek, J.A., and Clarke, N.D. (2005). Explicit equilibrium modeling of transcription-factor binding and gene regulation. Genome Biol **6,** R87.

Guertin, M.J., Martins, A.L., Siepel, A., and Lis, J.T. (2012). Accurate prediction of inducible transcription factor binding intensities in vivo. PLoS Genet **8,** e1002610.

Gumucio, D.L., Heilstedt-Williamson, H., Gray, T.A., Tarle, S.A., Shelton, D.A., Tagle, D.A., Slightom, J.L., Goodman, M., and Collins, F.S. (1992). Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. Mol Cell Biol **12,** 4919-4929.

Guo, C.L., Chen, L.G., He, X.H., Dai, Z., and Yuan, H.Y. (2005). [Expressions of LEAFY homologous genes in different organs and stages of Ginkgo biloba]. Yi Chuan **27,** 241-244.

Hamès, C., Ptchelkine, D., Grimm, C., Thevenon, E., Moyroud, E., Gerard, F., Martiel, J.L., Benlloch, R., Parcy, F., and Muller, C.W. (2008). Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. Embo J **27,** 2628-2637.

Hardison, R.C., and Taylor, J. (2012). Genomic approaches towards finding cis-regulatory modules in animals. Nat Rev Genet **13,** 469-483.

Himi, S., Sano, R., Nishiyama, T., Tanahashi, T., Kato, M., Ueda, K., and Hasebe, M. (2001). Evolution of MADS-box gene induction by FLO/LFY genes. J Mol Evol **53,** 387-393.

Hoekstra, H.E., and Coyne, J.A. (2007). The locus of evolution: evo devo and the genetics of adaptation. Evolution **61,** 995-1016.

Hofer, J., Turner, L., Hellens, R., Ambrose, M., Matthews, P., Michael, A., and Ellis, N. (1997). UNIFOLIATA regulates leaf and flower morphogenesis in pea. Curr Biol **7,** 581-587.

Hong, R.L., Hamaguchi, L., Busch, M.A., and Weigel, D. (2003). Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. Plant Cell **15,** 1296-1309.

Ikeda, K., Ito, M., Nagasawa, N., Kyozuka, J., and Nagato, Y. (2007). Rice ABERRANT PANICLE ORGANIZATION 1, encoding an F-box protein, regulates meristem fate. Plant J **51,** 1030-1040.

Immink, R.G., Kaufmann, K., and Angenent, G.C. (2010). The 'ABC' of MADS domain protein behaviour and interactions. Semin Cell Dev Biol **21,** 87-93.

John, S., Sabo, P.J., Thurman, R.E., Sung, M.H., Biddie, S.C., Johnson, T.A., Hager, G.L., and Stamatoyannopoulos, J.A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet **43,** 264-268.

Kafri, R., Springer, M., and Pilpel, Y. (2009). Genetic redundancy: new tricks for old genes. Cell **136,** 389-392.

Kanrar, S., Bhattacharya, M., Arthur, B., Courtier, J., and Smith, H.M. (2008). Regulatory networks that function to specify flower meristems require the function of homeobox genes PENNYWISE and POUND-FOOLISH in Arabidopsis. Plant J **54,** 924-937.

Kaufmann, K., Muino, J.M., Jauregui, R., Airoldi, C.A., Smaczniak, C., Krajewski, P., and Angenent, G.C. (2009). Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the Arabidopsis flower. PLoS Biol **7,** e1000090.

Kaufmann, K., Wellmer, F., Muino, J.M., Ferrier, T., Wuest, S.E., Kumar, V., Serrano-Mislata, A., Madueno, F., Krajewski, P., Meyerowitz, E.M., Angenent, G.C., and

Riechmann, J.L. (2010). Orchestration of floral initiation by APETALA1. Science **328,** 85-89.

Kitahara, K., Hibino, Y., Aida, R., and Matsumoto, S. (2004). Ectopic expression of the rose AGAMOUS-like MADS-box genes 'MASAKO C1 and D1' causes similar homeotic transformation of sepal and petal in Arabidopsis and sepal in Torenia. Plant Science **166,** 1245-1252.

Kuo, D., Licon, K., Bandyopadhyay, S., Chuang, R., Luo, C., Catalana, J., Ravasi, T., Tan, K., and Ideker, T. (2010). Coevolution within a transcriptional network by compensatory trans and cis mutations. Genome Res **20,** 1672-1678.

Kyozuka, J., Konishi, S., Nemoto, K., Izawa, T., and Shimamoto, K. (1998). Down-regulation of RFL, the FLO/LFY homolog of rice, accompanied with panicle branch initiation. Proc Natl Acad Sci U S A **95,** 1979-1982.

Lee, E.K., Cibrian-Jaramillo, A., Kolokotronis, S.O., Katari, M.S., Stamatakis, A., Ott, M., Chiu, J.C., Little, D.P., Stevenson, D.W., McCombie, W.R., Martienssen, R.A., Coruzzi, G., and Desalle, R. (2011). A functional phylogenomic view of the seed plants. PLoS Genet **7,** e1002411.

Lee, I., Wolfe, D.S., Nilsson, O., and Weigel, D. (1997). A LEAFY co-regulator encoded by UNUSUAL FLORAL ORGANS. Curr Biol **7,** 95-104.

Leitch, I.J., Hanson, L., Lim, K.Y., Kovarik, A., Chase, M.W., Clarkson, J.J., and Leitch, A.R. (2008). The ups and downs of genome size evolution in polyploid species of Nicotiana (Solanaceae). Ann Bot **101,** 805-814.

Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N., and Wasserman, W.W. (2003). Identification of conserved regulatory elements by comparative genome analysis. J Biol **2,** 13.

Levin, J.Z., and Meyerowitz, E.M. (1995). *UFO*: an Arabidopsis gene involved in both floral meristem and floral organ development. Plant Cell **7,** 529-548.

Lickwar, C.R., Mueller, F., Hanlon, S.E., McNally, J.G., and Lieb, J.D. (2012). Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. Nature **484,** 251-255.

Liljegren, S.J., Gustafson-Brown, C., Pinyopich, A., Ditta, G.S., and Yanofsky, M.F. (1999). Interactions among APETALA1, LEAFY, and TERMINAL FLOWER1 specify meristem fate. Plant Cell **11,** 1007-1018.

Liu, C., Thong, Z., and Yu, H. (2009a). Coming into bloom: the specification of floral meristems. Development **136,** 3379-3391.

Liu, C., Xi, W., Shen, L., Tan, C., and Yu, H. (2009b). Regulation of floral patterning by flowering time genes. Dev Cell **16,** 711-722.

Liu, X., and Clarke, N.D. (2002). Rationalization of gene regulation by a eukaryotic transcription factor: calculation of regulatory region occupancy from predicted binding affinities. J Mol Biol **323,** 1-8.

Lohmann, J.U., Hong, R.L., Hobe, M., Busch, M.A., Parcy, F., Simon, R., and Weigel, D. (2001). A molecular link between stem cell regulation and floral patterning in Arabidopsis. Cell **105,** 793-803.

Mahony, S., and Benos, P.V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res **35,** W253-258.

Maizel, A., Busch, M.A., Tanahashi, T., Perkovic, J., Kato, M., Hasebe, M., and Weigel, D. (2005). The floral regulator LEAFY evolves by substitutions in the DNA binding domain. Science **308,** 260-263.

Malek, O., Lattig, K., Hiesel, R., Brennicke, A., and Knoop, V. (1996). RNA editing in bryophytes and a molecular phylogeny of land plants. Embo J **15,** 1403-1411.

Man, T.K., and Stormo, G.D. (2001). Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. Nucleic Acids Res **29,** 2471-2478.

Mandel, M.A., and Yanofsky, M.F. (1995). A gene triggering flower development in Arabidopsis. Nature **377,** 522-524.

Mannervik, M., Nibu, Y., Zhang, H., and Levine, M. (1999). Transcriptional coregulators in development. Science **284,** 606-609.

Mellerowicz, E.J., Horgan, K., Walden, A., Coker, A., and Walter, C. (1998). PRFLL--a Pinus radiata homologue of FLORICAULA and LEAFY is expressed in buds containing vegetative shoot and undifferentiated male cone primordia. Planta **206,** 619-629.

Melzer, R., and Theissen, G. (2009). Reconstitution of 'floral quartets' in vitro involving class B and class E floral homeotic proteins. Nucleic Acids Res **37,** 2723-2736.

Meyer, M., Stenzel, U., Myles, S., Prufer, K., and Hofreiter, M. (2007). Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Res **35,** e97.

Moore, M.J., Bell, C.D., Soltis, P.S., and Soltis, D.E. (2007). Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci U S A **104,** 19363-19368.

Mouradov, A., Glassick, T., Hamdorf, B., Murphy, L., Fowler, B., Marla, S., and Teasdale, R.D. (1998). NEEDLY, a Pinus radiata ortholog of FLORICAULA/LEAFY genes, expressed in both reproductive and vegetative meristems. Proc Natl Acad Sci U S A **95,** 6537-6542.

Moyroud, E., Reymond, M.C., Hames, C., Parcy, F., and Scutt, C.P. (2009). The analysis of entire gene promoters by surface plasmon resonance. Plant J **59,** 851-858.

Moyroud, E., Kusters, E., Monniaux, M., Koes, R., and Parcy, F. (2010). LEAFY blossoms. Trends Plant Sci **15,** 346-352.

Moyroud, E., Minguet, E.G., Ott, F., Yant, L., Pose, D., Monniaux, M., Blanchet, S., Bastien, O., Thevenon, E., Weigel, D., Schmid, M., and Parcy, F. (2011). Prediction of Regulatory Interactions from Genome Sequences Using a Biophysical Model for the Arabidopsis LEAFY Transcription Factor. Plant Cell.

Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D., Snyder, M., Young, R.A., and Bulyk, M.L. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat Genet **36,** 1331-1339.

Muller, G.B. (2007). Evo-devo: extending the evolutionary synthesis. Nat Rev Genet **8,** 943-949.

Nishiyama, T., and Kato, M. (1999). Molecular phylogenetic analysis among bryophytes and tracheophytes based on combined data of plastid coded genes and the 18S rRNA gene. Mol Biol Evol **16,** 1027-1036.

Parcy, F. (2005). Flowering: a time for integration. Int J Dev Biol **49,** 585-593.

Parcy, F., Nilsson, O., Busch, M.A., Lee, I., and Weigel, D. (1998). A genetic framework for floral patterning. Nature **395,** 561-566.

Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet **10,** 669-680.

Pennisi, E. (2008). Evolutionary biology. Deciphering the genetics of evolution. Science **321,** 760-763.

Pinyopich, A., Ditta, G.S., Savidge, B., Liljegren, S.J., Baumann, E., Wisman, E., and Yanofsky, M.F. (2003). Assessing the redundancy of MADS-box genes during carpel and ovule development. Nature **424,** 85-88.

Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y., and Pritchard, J.K. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res **21,** 447-455.

Posé, D., Yant, L., and Schmid, M. (2012). The end of innocence: flowering networks explode in complexity. Curr Opin Plant Biol **15,** 45-50.

Prigge, M.J., and Bezanilla, M. (2010). Evolutionary crossroads in developmental biology: Physcomitrella patens. Development **137,** 3535-3543.

Qiu, Y.-L. (2008a). Phylogeny and evolution of charophytic algae and land plants. Journal of Systematics and Evolution **46,** 287-306.

Qiu, Y.L. (2008b). Phylogeny and evolution of charophytic algae and land plants. Journal of Systematics and Evolution **46,** 287-306.

Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E.D. (2002). Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. BMC Bioinformatics **3,** 30.

Rao, N.N., Prasad, K., Kumar, P.R., and Vijayraghavan, U. (2008). Distinct regulatory role for RFL, the rice LFY homolog, in determining flowering time and plant architecture. Proc Natl Acad Sci U S A **105,** 3646-3651.

Ratcliffe, O.J., Bradley, D.J., and Coen, E.S. (1999). Separation of shoot and floral identity in Arabidopsis. Development **126,** 1109-1120.

Rensing, S.A., Ick, J., Fawcett, J.A., Lang, D., Zimmer, A., Van de Peer, Y., and Reski, R. (2007). An ancient genome duplication contributed to the abundance of metabolic genes in the moss Physcomitrella patens. BMC Evol Biol **7,** 130.

Revet, B., von Wilcken-Bergmann, B., Bessert, H., Barker, A., and Muller-Hill, B. (1999). Four dimers of lambda repressor bound to two suitably spaced pairs of lambda operators form octamers and DNA loops over large distances. Curr Biol **9,** 151-154.

Roider, H.G., Kanhere, A., Manke, T., and Vingron, M. (2007). Predicting transcription factor affinities to DNA from a biophysical model. Bioinformatics **23,** 134-141.

Rokas, A. (2008). The molecular origins of multicellular transitions. Curr Opin Genet Dev **18,** 472-478.

Ronshaugen, M., McGinnis, N., and McGinnis, W. (2002). Hox protein mutation and macroevolution of the insect body plan. Nature **415,** 914-917.

Saddic, L.A., Huvermann, B., Bezhani, S., Su, Y., Winter, C.M., Kwon, C.S., Collum, R.P., and Wagner, D. (2006). The LEAFY target LMI1 is a meristem identity regulator and acts together with LEAFY to regulate expression of CAULIFLOWER. Development **133,** 1673-1682.

Saidi, Y., Finka, A., Chakhporanian, M., Zryd, J.P., Schaefer, D.G., and Goloubinoff, P. (2005). Controlled expression of recombinant proteins in Physcomitrella patens by a conditional heat-shock promoter: a tool for plant research and biotechnology. Plant Mol Biol **59,** 697-711.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res **32,** D91-94.

Schaefer, D.G. (2001). Gene targeting in Physcomitrella patens. Curr Opin Plant Biol **4,** 143-150.

Schmid, M., Uhlenhaut, N.H., Godard, F., Demar, M., Bressan, R., Weigel, D., and Lohmann, J.U. (2003). Dissection of floral induction pathways using global expression analysis. Development **130,** 6001-6012.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D., and Lohmann, J.U. (2005). A gene expression map of Arabidopsis thaliana development. Nat Genet **37,** 501-506.

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., Talianidis, I., Flicek, P., and Odom, D.T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science **328,** 1036-1040.

Schneider, T.D. (2002). Consensus sequence Zen. Appl Bioinformatics **1,** 111-119.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res **18,** 6097-6100.

Schultz, E.A., and Haughn, G.W. (1991). *LEAFY*, a homeotic gene that regulates inflorescence development in Arabidopsis. Plant Cell **3,** 771-781.

Schwarz-Sommer, Z., Hue, I., Huijser, P., Flor, P.J., Hansen, R., Tetens, F., Lönnig, W.-E., Saedler, H., and Sommer, H. (1992). Characterization of the *Antirrhinum* floral homeotic MADS-box gene *deficiens*: evidence for DNA binding and autoregulation of its persistent expression throughout flower development. EMBO Journal. **11,** 251-263.

Segal, E., and Widom, J. (2009). What controls nucleosome positions? Trends Genet **25,** 335-343.

Shannon, S., and Meek-Wagner, D.R. (1993). Genetic Interactions That Regulate Inflorescence Development in Arabidopsis. Plant Cell **5,** 639-655.

Shiokawa, T., Yamada, S., Futamura, N., Osanai, K., Murasugi, D., Shinohara, K., Kawai, S., Morohoshi, N., Katayama, Y., and Kajita, S. (2008). Isolation and functional analysis of the CjNdly gene, a homolog in Cryptomeria japonica of FLORICAULA/LEAFY genes. Tree Physiol **28,** 21-28.

Shiu, S.H., Shih, M.C., and Li, W.H. (2005). Transcription factor families have much higher expansion rates in plants than in animals. Plant Physiol **139,** 18-26.

Singer, S.D., Krogan, N.T., and Ashton, N.W. (2007). Clues about the ancestral roles of plant MADS-box genes from a functional analysis of moss homologues. Plant Cell Rep **26,** 1155-1169.

Sinha, S., Schroeder, M.D., Unnerstall, U., Gaul, U., and Siggia, E.D. (2004). Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. BMC Bioinformatics **5,** 129.

Siriwardana, N.S., and Lamb, R.S. (2012). The poetry of reproduction: the role of LEAFY in Arabidopsis thaliana flower formation. Int J Dev Biol **56,** 207-221.

Smyth, D.R., Bowman, J.L., and Meyerowitz, E.M. (1990). Early flower development in Arabidopsis. Plant Cell **2,** 755-767.

Soltis, D.E., and Soltis, P.S. (2004). Amborella not a "basal angiosperm"? Not so fast. Am J Bot **91,** 997-1001.

Soltis, D.E., Bell, C.D., Kim, S., and Soltis, P.S. (2008). Origin and early evolution of angiosperms. Ann N Y Acad Sci **1133,** 3-25.

Southerton, S.G., Strauss, S.H., Olive, M.R., Harcourt, R.L., Decroocq, V., Zhu, X., Llewellyn, D.J., Peacock, W.J., and Dennis, E.S. (1998). Eucalyptus has a functional equivalent of the Arabidopsis floral meristem identity gene LEAFY. Plant Mol Biol **37,** 897-910.

Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., Parts, L., Carlson, J.W., Crosby, M.A., Rasmussen, M.D., Roy, S., Deoras, A.N., Ruby, J.G., Brennecke, J., Hodges, E., Hinrichs, A.S., Caspi, A., Paten, B., Park, S.W., Han, M.V., Maeder, M.L., Polansky, B.J., Robson, B.E., Aerts, S., van Helden, J., Hassan, B., Gilbert, D.G., Eastman, D.A., Rice, M., Weir, M., Hahn, M.W., Park, Y., Dewey, C.N., Pachter, L., Kent, W.J., Haussler, D., Lai, E.C., Bartel, D.P., Hannon, G.J., Kaufman, T.C., Eisen, M.B., Clark, A.G., Smith, D., Celniker, S.E., Gelbart, W.M., and Kellis, M. (2007). Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature **450,** 219-232.

Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics **16,** 16-23.

Tanahashi, T., Sumikawa, N., Kato, M., and Hasebe, M. (2005). Diversification of gene function: homologs of the floral regulator FLO/LFY control the first zygotic cell division in the moss Physcomitrella patens. Development **132,** 1727-1736.

Timme, R.E., Bachvaroff, T.R., and Delwiche, C.F. (2012). Broad phylogenomic sampling and the sister lineage of land plants. PLoS One **7,** e29696.

Trouiller, B., Schaefer, D.G., Charlot, F., and Nogue, F. (2006). MSH2 is essential for the preservation of genome integrity and prevents homeologous recombination in the moss Physcomitrella patens. Nucleic Acids Res **34,** 232-242.

Tsong, A.E., Tuch, B.B., Li, H., and Johnson, A.D. (2006). Evolution of alternative transcriptional circuits with identical logic. Nature **443,** 415-420.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science **249,** 505-510.

Vazquez-Lobo, A., Carlsbecker, A., Vergara-Silva, F., Alvarez-Buylla, E.R., Pinero, D., and Engstrom, P. (2007). Characterization of the expression patterns of LEAFY/FLORICAULA and NEEDLY orthologs in female and male cones of the conifer genera Picea, Podocarpus, and Taxus: implications for current evo-devo hypotheses for gymnosperms. Evol Dev **9,** 446-459.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet **5,** 276-287.

Wei, G.H., Badis, G., Berger, M.F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A.R., Yan, J., Talukder, S., Turunen, M., Taipale, M., Stunnenberg, H.G., Ukkonen, E., Hughes, T.R., Bulyk, M.L., and Taipale, J. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. Embo J **29,** 2147-2160.

Weigel, D., and Nilsson, O. (1995). A developmental switch sufficient for flower initiation in diverse plants. Nature **377,** 495-500.

Weigel, D., Alvarez, J., Smyth, D.R., Yanofsky, M.F., and Meyerowitz, E.M. (1992). LEAFY controls floral meristem identity in Arabidopsis. Cell **69,** 843-859.

Wellik, D.M. (2009). Hox genes and vertebrate axial pattern. Curr Top Dev Biol **88,** 257-278.

William, D.A., Su, Y., Smith, M.R., Lu, M., Baldwin, D.A., and Wagner, D. (2004). Genomic identification of direct target genes of LEAFY. Proc Natl Acad Sci U S A **101,** 1775-1780.

Winter, C.M., Austin, R.S., Blanvillain-Baufume, S., Reback, M.A., Monniaux, M., Wu, M.F., Sang, Y., Yamaguchi, A., Yamaguchi, N., Parker, J.E., Parcy, F., Jensen, S.T., Li, H., and Wagner, D. (2011). LEAFY target genes reveal floral regulatory logic, cis motifs, and a link to biotic stimulus response. Dev Cell **20,** 430-443.

Workman, C.T., Yin, Y., Corcoran, D.L., Ideker, T., Stormo, G.D., and Benos, P.V. (2005). enoLOGOS: a versatile web tool for energy normalized sequence logos. Nucleic Acids Res **33,** W389-392.

Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., and Romano, L.A. (2003). The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol **20,** 1377-1419.

Wu, X., Dinneny, J.R., Crawford, K.M., Rhee, Y., Citovsky, V., Zambryski, P.C., and Weigel, D. (2003). Modes of intercellular transcription factor movement in the Arabidopsis apex. Development **130,** 3735-3745.

Zhang, W., Zhang, T., Wu, Y., and Jiang, J. (2012). Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. Plant Cell **24,** 2719-2731.

Zobell, O., Faigl, W., Saedler, H., and Munster, T. (2010). MIKC* MADS-box proteins: conserved regulators of the gametophytic generation of land plants. Mol Biol Evol **27,** 1201-1211.

Résumé

LEAFY (LFY) est un facteur de transcription unique et très conservé chez les plantes terrestres. Il contrôle le développement floral chez les angiospermes (plantes à fleurs), mais son rôle est encore mal connu chez toutes les autres plantes terrestres à l'exception de la mousse *Physcomitrella patens* où l'orthologue de LFY (PpLFY) est requis pour la première division cellulaire du zygote. PpLFY ne reconnaît pas les mêmes séquences d'ADN que LFY d'*Arabidopsis thaliana*, malgré la très forte conservation de leurs domaines de liaison à l'ADN. LFY semble donc avoir changé de propriétés au cours de l'évolution ; l'objectif de ma thèse a été de déterminer si de tels changements s'étaient produits fréquemment chez les plantes terrestres, et de comprendre leur origine et leur impact sur la régulation des gènes cibles de LFY.
Pour cela, j'ai étudié la spécificité de liaison à l'ADN des orthologues de LFY chez les grands groupes de plantes terrestres par des expériences de SELEX, et cette spécificité s'est révélée très fortement conservée, excepté dans le cas de PpLFY. Ces résultats nous ont permis de construire un modèle biophysique performant pour prédire la liaison de LFY à l'échelle génomique, ce que nous avons appliqué à l'étude de l'évolution de la régulation de quelques gènes clés par LFY. Nous avons ainsi pu prédire la régulation du gène floral *AGAMOUS* par LFY chez différentes espèces angiospermes, et nous avons pu montrer que LFY régulait très vraisemblablement les orthologues des gènes d'identité florale chez les gymnospermes, c'est-à-dire avant l'apparition de la fleur.
La divergence de spécificité de PpLFY nous a poussés à étudier les gènes cibles de PpLFY : pour cela, j'ai initié des approches bioinformatiques et expérimentales chez *P. patens*. Enfin, pour comprendre comment ce changement de spécificité s'est déroulé au cours de l'évolution, nous nous sommes penchés sur l'ancêtre de LFY et avons découvert que LFY était déjà présent chez les algues vertes. Des études pour déterminer la spécificité ancestrale de LFY chez ces espèces ont été initiées.

Abstract

LEAFY (LFY) is a unique transcription factor, highly conserved within land plants. LFY directly regulates a set of genes participating in floral development in angiosperms (flowering plants), but its role in the other groups of land plants is unknown, except in the moss *Physcomitrella patens* where the LFY ortholog (PpLFY) regulates the first cell division in the zygote. PpLFY does not bind to the same DNA sequences as LFY from *Arabidopsis thaliana*, in spite of the very high degree of conservation of their DNA binding domains. Thus, it appears that the properties of LFY have changed during evolution ; the goal of my thesis was to find out if such changes had occurred frequently in land plants, and what are their origins and consequences on target genes regulation.
I performed SELEX experiments on LFY orthologs from all land plants, which revealed that their DNA binding specificty was highly conserved, except in the case of PpLFY. These results allowed us to build an accurate biophysical model to predict LFY binding on DNA fragments at a genomic level, which we applied on the evolution of the regulation of key target genes by LFY. We were able to predict the regulation of the floral gene *AGAMOUS* by LFY in various angiosperm species, et we could also show that LFY was very likely regulating gymnosperm orthologs of genes involved in floral organ identity, even before the appearance of the flower.
The change in DNA binding specificity observed for PpLFY led us to study more precisely the consequences of this change for the regulation of target genes : for this, I initiated bioinformatic and experimental work in *P. patens*. Finally, to understand how this change in DNA binding specificity had occurred during evolution, we looked for the ancestor of LFY and found out that LFY already existed in green algae. We are currently investigating the ancestral specificity of LFY in these species.